ABSTRACT

Title of dissertation: A PROGRAM FOR EXPERIMENTAL SYNTAX: FINDING THE RELATIONSHIP BETWEEN ACCEPTABILITY AND GRAMMATICAL KNOWLEDGE

Jon Sprouse
Doctor of Philosophy, 2007

Dissertation directed by: Professor Howard Lasnik
Department of Linguistics

There has always been interest in the methodology of acceptability judgment collection, as well as the reliability of the results. It seems, though, that the past several years have seen an increase in the number of studies employing formal experimental techniques for the collection of acceptability judgments, so much so that the term *experimental syntax* has come to be applied to the use of those techniques. The question this dissertation asks is whether the extent of the utility of experimental syntax is to find areas in which informal judgment collection was insufficient, or whether there is a complementary research program for experimental syntax that is more than just a methodological footnote to the informal judgment collection of theoretical syntax. This dissertation is a first attempt at a tentative *yes*: the tools of experimental syntax can be used to explore the relationship between acceptability judgments and the form or nature of grammatical knowledge, not just the content of grammatical knowledge. This dissertation begins by identifying several recent claims about the nature of grammatical knowledge that have been made based upon hypotheses about the

nature of acceptability judgments. Each chapter applies the tools of experimental syntax to those hypotheses in an attempt to refine our understanding of the relationship between acceptability and grammatical knowledge. The claims investigated include: that grammatical knowledge is gradient, that grammatical knowledge is sensitive to context effects, that the stability or instability of acceptability reflects underlying differences in grammatical knowledge, that processing effects affect acceptability, and that acceptability judgments have nothing further to contribute to debates over the number and nature of dependency forming operations. Using wh-movement and Island effects as the empirical basis of the research, the results of these studies suggest that the relationship between acceptability and grammatical knowledge is much more complicated than previously thought. The overarching conclusion is that there is a program for experimental syntax that is independent of simple data collection: only through the tools of experimental syntax can we achieve a better understanding of the nature of acceptability, and how it relates to the nature of grammatical knowledge.

A Program for Experimental Syntax: Finding the relationship between acceptability and grammatical knowledge

by

Jon Sprouse

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Howard Lasnik, Chair
Professor Norbert Hornstein
Associate Professor Jeffrey Lidz
Associate Professor Colin Phillips
Professor Nan Ratner

ACKNOWLEDGMENTS

Words cannot express the gratitude, and appreciation, I feel for all of the wonderful people in my life, but this page will have to be a feeble attempt. I will apologize now for any sentiments I forget to express, or any names I forget to mention. Luckily, if you have been an important part of my life and are reading this dissertation, you already know that I love you.

First and foremost, I owe more than I can say to my advisor Howard Lasnik. I was once told by a close friend that Howard is the second greatest living syntactician. I am not qualified to judge that statement (although anyone who has read Howard's work can attest to his brilliance), but I can say this: He is by far one of the greatest living advisors, both within the field of linguistics and without, and the best teacher I could ever have hoped to have.

Although he would never admit it, Norbert Hornstein also deserves a large debt of gratitude for guiding me through graduate school. Always content to let others take the credit, Norbert is a de facto advisor to all of the students in the department, and in many ways a role model for all of us. He has that rare ability to find the best in everyone, and coax it out (by force if necessary). I can only hope to one day have as much faith in myself as he has in me.

Robert Freidin also deserves the credit (or blame) for pulling me into the field of linguistics. He was willing to invest his time in an ignorant undergraduate, and

for that I will always be grateful. While it would have been easy enough for him to stop advising me after I moved on to graduate school, he has always been willing to offer advice and encouragement, accepting me as a colleague despite the fact that I am nowhere near his equal. Though I can never thank him appropriately for this, I can promise to remember his example while teaching my own undergraduate students one day.

It goes without saying that I am eternally grateful to my graduate school friends. I won't name them all here (you know who you are), but from conversations to gym trips, smoothie runs to Moe's burritos, it has been an incredible four years. And although this dissertation signifies my departure from you all too soon, you should know that it wouldn't exist without you.

And of course, I must thank my parents. Although they still can't understand why I have gone to school for this long and am not yet a millionaire, they have always been willing to let me follow my own path. I've learned in recent years that such freedom is a true rarity. I can only hope to remember that if I ever have children of my own.

Finally, to my girlfriend-fiancée-soon-to-be-wife, Caroline: You know that I am bad with words, and I fear that I only have one good attempt at expressing my love for you in me, so I will save that for the wedding. However, you should know that I could never - *never* - have finished this, or anything, without you. You are my best friend and the best thing that has ever happened to me. I can only hope that you will continue to happen to me for years to come.

Table of Contents

List of Tables vi

List of Figures ix

1 Introduction 1

2 A critical look at magnitude estimation 10
2.1 Magnitude estimation basics . . . 11
2.1.1 Magnitude estimation in psychophysics . . . 11
2.1.2 Magnitude estimation in linguistics . . . 16
2.2 The interval scale of linguistic magnitude estimation . . . 22
2.2.1 Internal consistency and cross-modality matching . . . 22
2.2.2 Interval regularity and the effect of the reference sentence . . . 24
2.2.3 Interval irregularity and its effect on theoretical syntax . . . 27
2.3 The distribution of acceptability judgments . . . 32
2.3.1 The log transformation . . . 32
2.3.2 The distribution of judgments in this dissertation . . . 36
2.3.2.1 Multiple observations per participant . . . 37
2.3.2.2 Few observations per participant, large sample . . . 38
2.3.2.3 Few observations per participant, small sample . . . 39
2.3.2.4 The overall distribution of judgments . . . 40
2.3.3 The effect of the log transformation on theoretical syntax . . . 41
2.4 The real benefits of magnitude estimation . . . 42

3 Context, acceptability, and grammaticality 46
3.1 The complexities of context . . . 47
3.2 Island effects and knowledge of the intended meaning . . . 52
3.2.1 Results . . . 59
3.3 Conflicting attention and non-Island unacceptability . . . 63
3.4 Discourse Linking and context . . . 69
3.4.1 The Superiority effect and D-linking . . . 69
3.4.2 The Resumption effect and D-linking . . . 71
3.4.3 The D-linking experiment . . . 73
3.5 The complexity of context and wh-movement . . . 80

4 Satiation and types of grammatical knowledge 81
4.1 The problem of *Syntactic Satiation* . . . 82
4.2 The replication problem . . . 85
4.2.1 Confirming the replication problem . . . 85
4.2.2 Deconstructing Snyder 2000 . . . 90
4.2.3 A roadmap . . . 96
4.3 The yes/no task and balanced/unbalanced designs . . . 97
4.4 The magnitude estimation task and balanced/unbalanced designs . . . 100

4.4.1 Magnitude estimation and balanced designs 101
4.4.2 The memory confound 109
4.4.3 Magnitude estimation and unbalanced designs 111
4.5 Conclusion 116
4.6 Some remaining questions 118
4.6.1 Whether Islands versus the That-trace effect 118
4.6.2 What about the models of satiation? 120

5 Processing effects and the differential sensitivity of acceptability 122
5.1 Syntactically and semantically ungrammatical representations 123
5.1.1 Some psycholinguistic background 123
5.1.2 Rationale 125
5.2 The reanalysis confound 132
5.3 The differential sensitivity of acceptability to processing effects 136

6 The role of acceptability in theories of wh-in-situ 137
6.1 Island effects and wh-in-situ in English 139
6.1.1 Testing Huang 1982 140
6.1.2 A better design for investigating covert movement Islands . . . 144
6.1.3 Categorical acceptability of overt and covert movement Islands 149
6.1.4 Theoretical implications of Subject Island effects with wh-in-situ 151
6.2 Distance effects and wh-in-situ in English 153
6.2.1 Distance and wh-dependencies 154
6.2.2 Distance and Binding dependencies 159
6.3 Refining the analyses of wh-in-situ in English 163

7 Conclusion 166

List of Tables

2.1 Characteristic exponents for various physical stimuli 16

2.2 Reference sentences for each experiment 24

2.3 Conditions in both experiments 25

2.4 Conditions means for If-reference 25

2.5 Predicted means for CSC-reference based on If-reference results . . . 26

2.6 Predicted and actual means for CSC-reference 26

2.7 Paired samples t-test results for both experiments 30

2.8 Qualitative normality results for multiple observations per participant 37

2.9 Qualitative normality results for few observations per participant, large sample size 39

2.10 Qualitative normality results for few observations per participant, small sample size 40

2.11 Qualitative normality results for multiple observations per participant 40

3.1 Islands: descriptive results without context 59

3.2 Islands: descriptive results with context 59

3.3 Results for three-way repeated measures ANOVA 60

3.4 Partial Eta-squared effect sizes (Cohen 1973) 60

3.5 Pairs from Deane 1991 65

3.6 Mean values and standard deviations for each condition 67

3.7 Two-way repeated measures ANOVAs using ordinal and rank data . . 68

3.8 Superiority: descriptive results 76

3.9 Resumption: descriptive results 76

3.10 Results for three-way repeated measures ANOVA, untransformed data 76

3.11 Results for three-way repeated measures ANOVA, rank-transformed data 77

4.1 Violations tested in Snyder 2000 83
4.2 Summary of results for Snyder 2000, Hiramatsu 2000, and Goodall 2005 84
4.3 Violations used in the modified replication attempt 88
4.4 Results from the direct replication 89
4.5 Results from the replication with confidence 89
4.6 Results from the modified replication 89
4.7 Summary of Snyder 2000 and 5 subsequent replication attempts . . . 90
4.8 Crossed design of factors Task and Design 96
4.9 Design manipulation for yes/no task 97
4.10 Results from the unbalanced yes/no task 99
4.11 Results from the balanced yes/no task 99
4.12 Composition of each block in balanced magnitude estimation experiments 103
4.13 Exemplars of each of the Islands investigated 104
4.14 Linear regressions for means of magnitude estimation in a balanced design 107
4.15 Linear regressions for residuals of magnitude estimation in a balanced design 108
4.16 Linear regressions for means after removing participants 110
4.17 Violations in unbalanced MagE task 113
4.18 Linear regressions for means, CSC-reference 113
4.19 Linear regressions for means, If-reference 114
4.20 Linear regressions for residuals, CSC-reference 114
4.21 Linear regressions for residuals, If-reference 115
4.22 Crossed design of factors Task and Design 117
4.23 Relative acceptability versus satiation from Snyder 2000 119
4.24 Relative acceptability versus satiation based on non-correctability . . 120

5.1 Reading times (in ms) at critical words for each dependency, from Stowe 1986 124

5.2 Reading times at critical words by filler type, from Pickering and Traxler 2003 125

5.3 Mean complexity ratings for short and long movement, from Phillips et al. 2005 127

5.4 Results and paired t-tests for experiment 1 130

5.5 Results for experiment 2 134

6.1 Results and paired t-tests for Huang-style conditions 142

6.2 Wh-in-situ: descriptive results 148

6.3 Wh-in-situ: Two-way repeated measures ANOVA 148

6.4 Categorical acceptability of overt movement and wh-in-situ in Islands 150

6.5 Mean complexity ratings for short and long movement, from Phillips et al. 2005 154

6.6 Paired t-tests for wh-movement distance conditions 157

6.7 Results for binding distance t-tests 162

List of Figures

2.1 Relative means for both experiments 27

2.2 Pattern of acceptability for both ratios and logs 30

3.1 Two main effects, no interaction, no Island effect 54

3.2 Interaction, Island effect 54

3.3 Island effects and context 61

3.4 Superiority and D-Linking 78

3.5 Resumption and D-Linking 80

4.1 Three models of satiation 94

4.2 Scatterplots and trendlines for Subject Islands 108

6.1 Overt versus covert Island effects following Huang 1982 143

6.2 Effects for wh-movement distance 157

6.3 Results for binding distance 161

Chapter 1

Introduction

As the extensive review by Schütze 1996 demonstrates, there has always been interest in the methodology of acceptability judgment collection, as well as the reliability of the results. It seems, though, that the past several years have seen an increase in the number of studies employing formal experimental techniques for the collection of acceptability judgments, so much so that the term *experimental syntax* has come to be applied to the use of those techniques (perhaps from the title of Cowart 1997). Indeed, there have been several journal articles in recent years advocating the use of such techniques by demonstrating the new data that they may reveal, and how this new data prompts reconsideration of theoretical analyses. The question this dissertation asks is whether that is the extent of the utility of experimental syntax - to find areas in which informal judgment collection was insufficient - or whether there is a complementary research program for experimental syntax that is more than just a methodological footnote to the informal judgment collection of theoretical syntax. This dissertation is a first attempt at a tentative *yes*: the tools of experimental syntax can be used to explore the relationship between acceptability judgments and the form or nature of grammatical knowledge, not just the content of grammatical knowledge.

In many ways, the driving questions of theoretical syntax can be summarized by the following two questions:

1. What is the content of grammatical knowledge?

2. What is the nature of grammatical knowledge?

Potential answers to the first question arise in any theoretical syntax paper: Island constraint X is or is not part of language Y, binding principle W is or is not part of language Z, et cetera. The potential answers to the second question are much more complicated, and rarely appear in any single analysis. In fact, they usually form the foundation for the different approaches to syntactic theory: transformations are or are not necessary, dependency formation is or is not constrained by locality conditions, grammatical knowledge is binary or gradient, et cetera. It is straightforward to apply the tools of experimental syntax to the first question, as acceptability judgments form the primary data for most syntactic analyses. However, applying experimental syntax to the second question is less straightforward, as it is not always clear what role individual acceptability facts play in supporting these claims about grammatical knowledge.

This dissertation begins by identifying several recent claims about the nature of grammatical knowledge that have been made based upon hypotheses about the nature of acceptability judgments. Each chapter applies the tools of experimental syntax to those hypotheses in an attempt to refine our understanding of the relationship between acceptability and grammatical knowledge. While each chapter has its own theoretical implications, the overarching conclusion is that there is a program for experimental syntax that is independent of simple data collection: only through the tools of experimental syntax can we achieve a better understanding of the nature of

acceptability, and how it relates to the nature of grammatical knowledge.

In order to keep the empirical reach of the dissertation manageable, attention will be focused on wh-questions, and in particular, the locality restrictions on wh-question formation known as Island effects(Ross 1967). As will soon become obvious, there is no shortage of claims about the nature of the grammar that have been made on the basis of the acceptability of wh-questions. This is, of course, not entirely surprising given the amount of research that has been done on wh-questions over the past forty years. Each of the topics of the subsequent chapters are briefly summarized below.

Chapter 2: A critical look at magnitude estimation

Chapter 2 serves several interrelated purposes. First, chapter 2 provides an overview of both psychophysical and linguistic magnitude estimation for linguists who are unfamiliar the similarities and differences between the two techniques. Second, chapter 2 provides a critical assessment of the claim that magnitude estimation provides 'better' data than other acceptability collection techniques, focusing on the assertion that magnitude estimation provides interval level acceptability data. The results of that assessment suggest that linguistic magnitude estimation does not provide interval level data, and in fact, is more like the standard ordinal rating tasks than originally thought.

The final contribution of chapter 2 is to evaluate the claim that grammatical knowledge is gradient. The gradience claim is by no means new, but has received added attention over the past few years thanks in no small part to magnitude es-

timation and the continuous data that it yields (see especially Keller 2000, 2003, Sorace and Keller 2005). While obtaining gradient data from a continuous task such as magnitude estimation is not surprising, chapter 2 reports some surprising results that suggest that participants impose categorical judgments akin to 'grammatical' and 'ungrammatical' on the magnitude estimation task.

Chapter 3: Context, acceptability, and grammaticality

One of the recurrent questions facing syntactic research is whether the fact that acceptability judgments are given for isolated sentences out of their natural linguistic context affects the results. This chapter lays out several possible ways in which the presence or absence of context may affect the results of acceptability studies, and what effect each could have on theories of grammatical knowledge. Because the effect of context can only be determined empirically for individual grammatical principles on a case by case basis, this chapter then presents 3 case studies of the effect of various types of context on properties of wh-movement in an attempt to determine i) if there is an effect, and ii) if so, what the underlying source of the effect may be. These case studies serve a dual purpose: they serve as a first attempt at systematically investigating the effect of context on properties of wh-movement, and, given that wh-movement serves as the empirical object of study in the rest of the dissertation, they serve as a set of control experiments to determine whether context should be included as a factor in experiments in subsequent chapters. These experiments suggest that at least the types of context studied here do not affect major properties of wh-movement such as Island effects and D-linking effects, indicating that context need not be in-

cluded as a factor in subsequent studies of Island effects in the rest of the dissertation.

Chapter 4: Satiation and types of grammatical knowledge

This chapter investigates a phenomenon, known in the literature as *syntactic satiation*, in which acceptability judgments of certain violations appear to get better, that is more acceptable, after several repetitions. The fact that some violations satiate while others do not has been interpreted in the literature as an indication of different underlying sources of the unacceptability. In other words, the nature of the violation (or the nature of the grammatical knowledge) affects its relationship with acceptability such that acceptability may change over time.

This chapter presents evidence that the reported satiation data is not robust, in that it is not replicable. Several experiments are presented that attempt to track down the reason for the replication problem, in the process teasing apart the various factors that may contribute to the satiation effect. In the end, the results suggest that satiation is actually an artifact of experimental factors, not a true property of either acceptability judgments or grammatical constraints, thus eliminating the possibility of diagnosing types of grammatical knowledge based on satiation.

Chapter 5: Processing effects and the differential sensitivity of acceptability

Linguists have agreed since at least Chomsky 1965 that acceptability judgments are too coarse grained to distinguish between effects of grammatical knowledge (what Chomsky 1965 would call competence effects) and effects of implementing that knowledge (or performance effects). With the rise of experimental methodologies for

collecting acceptability judgments, there has been a renewed interest in attempting to identify the contribution of performance factors, in particular processing factors, to acceptability judgments. For instance, Fanselow and Frisch 2004 report that local ambiguity in German can lead to increases in acceptability, suggesting that the momentary possibility of two representation can affect acceptability. Sag et al. submitted report that factors affecting the acceptability of Superiority violations also affect the processing of wh-questions as measured in reading times, suggesting that there might be a correlation between processing factors and the acceptability of Superiority violations. This chapter builds on this work by investigating three different types of temporary representations created by the Active Filling processing strategy (Frazier and Flores d'Arcais 1989) to determine if they affect the acceptability of the final representation. The question is whether every type of processing effect that arises due to the active filling strategy affects acceptability, or whether acceptability is differentially sensitive to such processing effects. The results suggest that judgment tasks are indeed differentially sensitive: they are sensitive to some processing effects, but not others. This differential sensitivity in turn suggests that further research is required to determine the class of processing effects that affect acceptability in order to refine the relationship between acceptability, grammaticality, and processing effects. Determining that relationship will be the first step toward assessing the merits of both processing-based and grammatical knowledge-based explanations of acceptability.

Chapter 6: The role of acceptability in theories of wh-in-situ

While previous chapters in this dissertation focused on the relationship between acceptability and grammaticallity in an attempt to refine our understanding of the nature of grammatical knowledge, this chapter demonstrates a more direct relationship between understanding the source of acceptability and syntactic theories. The claim in this chapter is straightforward: there are new types of acceptability data that can be revealed through experimentation, and a better understanding of the relationship between these new acceptability effects and grammatical knowledge can have significant consequences for the set of possible dependency forming operations. This chapter focuses almost exclusively on wh-in-situ in multiple wh-questions in English. In-situ wh-words must be interpreted somehow, and that interpretation is dependent on the interpretation of the wh-word in matrix C. There is a general consensus that there must be a dependency between these two positions, but there is significant debate over the nature of that dependency, and in particular, over the dependency forming operation(s) that create it. Various proposals have been made in the literature, such as covert wh-movement (Huang 1982), null operator movement and unselective binding (Tsai 1994), choice-function application and existential closure (Reinhart 1997), overt movement and pronunciation of the lower copy (Bošković 2002), and long distance AGREE (Chomsky 2000). This chapter uses the two new pieces of evidence gained from experimental syntax to refine our understanding of the number and type of dependency forming operations for wh-in-situ dependencies in English.

The results of these investigations suggest that:

1. While grammatical knowledge may still be gradient, the categorical distinction between grammatical and ungrammatical is also real, as it arises (unprompted) even in continuous tasks like magnitude estimation .

2. While context may have an effect on some acceptability judgments, it is likely that those effects reflect non-structural constraints such as pragmatics or information structure. Furthermore, there is no evidence that context affects the wh-movement properties investigated in this dissertation: Island effects.

3. While there still may be different causes underlying various violations, satiation is not likely to differentiate among them, as satiation is most likely an artifact of experimental design, and not a property of acceptability judgments.

4. While it goes without saying that processing effects affect acceptability judgments, it is not the case that *all* processing effects have an effect. This differential sensitivity suggests a methodology for validating the possibility of processing-based explanations of acceptability effects: first investigate whether the processing effects in question have an effect on acceptability.

5. While the value of non-acceptability data such as possible answers are undoubtedly valuable for investigations of wh-in-situ, two new acceptability-based effects were presented that may have important consequences for wh-in-situ theories. While some hypotheses were suggested, future research is necessary to identify the relationship between those effects and grammatical knowledge.

Like many studies, the work presented in this dissertation raises far more questions than it answers. However, it is clear from these results that there is a good deal of potential for experimental syntax to provide more than a simple fact-checking service for theoretical syntax: the tools of experimental syntax are in a unique position to further our understanding of the relationship between acceptability and grammatical knowledge, and ultimately, refine our theories of the nature of grammatical knowledge itself.

Chapter 2

A critical look at magnitude estimation

Although logically separable, in many ways the term experimental syntax has become synonymous with the use of the magnitude estimation task to collect acceptability judgments. As such, this chapter serves two purposes: i) to provide an overview and assessment of the magnitude estimation task as applied to acceptability judgments, and ii) to provide a foundation from which theoretical syntacticians can begin to evaluate the claim that magnitude estimation reveals the gradient nature of grammatical knowledge. After a brief comparison of magnitude estimation in psychophysics and linguistics, the claim that linguistic magnitude estimation data has the same properties as psychophysical magnitude estimation data will be investigated through meta analyses of several of the experiments that will be presented in the rest of the dissertation. The surprising finding from these analyses is that at least two of the mathematical properties attributed to linguistic magnitude estimation data are untrue (that the intervals between data points are regular, and that the responses are log-normally distributed), which in fact has led to an inappropriate methodology for analyzing the data. These results suggest that linguistic magnitude estimation is mathematically more similar to linguistic 5- or 7-point scale rating tasks than to psychophysical magnitude estimation. These findings are interpreted with respect to claims in the literature that magnitude estimation demonstrates the gradient nature

of linguistic knowledge, and suggestions are made as to what the true benefit of magnitude estimation may be: the freedom it gives participants to convey any distinction they see as relevant. In fact, surprising evidence emerges that participants use this freedom to convey a categorical distinction between grammatical and ungrammatical sentences, despite the lack of any explicit or implicit mention of a categorical grammaticality distinction in the magnitude estimation task.

2.1 Magnitude estimation basics

2.1.1 Magnitude estimation in psychophysics

Magnitude estimation was first employed by the field of psychophysics to study the relationship between the physical strength of a given stimulus, such as the brightness of light, and the perceived strength of that stimulus. This subsection provides a brief overview of magnitude estimation in psychophysics as a baseline for understanding, and evaluating, magnitude estimation of linguistic acceptability.

It is perhaps easiest to begin our discussion of magnitude estimation with an example. Imagine you are presented with a set of lines. The first line is the *modulus* or *reference*, and you are told its length is 100 units. You can use this information to estimate the length of the other lines using your perception of visual length. For instance, if you believe that the line labeled **Item 1** is twice as long as the reference line, you could assign it a length of 200 units. If you believe **Item 2** is half as long as the reference line, you could assign it a length of 50 units: The resulting data are estimates of the length of the items in units equal to the length of the reference

Reference: _____
Length: 100

Item 1: _____
Length: *200*

Item 2: _____
Length: *50*

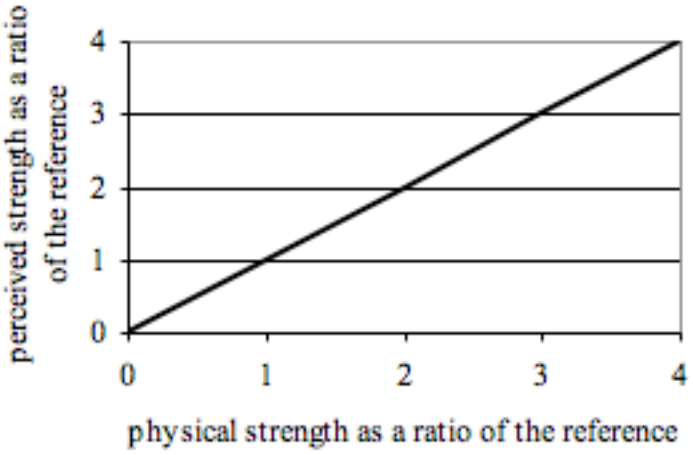
Item 3: _____
Length: *300*

line. These estimates can be compared to the actual physical length of the lines: the reference line is actually 100 points, or about 1.4 inches, item 1 is 200 points (2.8 inches), item 2 is 50 points (.7 inches) and item 3 is 300 points (4.2 inches). By comparing this physical measurement to the perceptual estimation for several different lengths of lines, psychologists can determine how accurate humans are at perceiving line length: Do humans overestimate or underestimate line length? Does the length of the line affect perception?

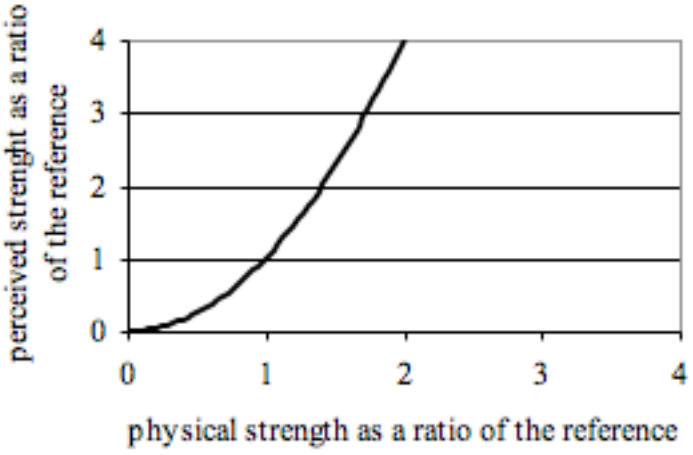
While line length is a simple case, the magnitude estimation technique can be extended to any physical stimulus, and indeed, over the past 50 years magnitude estimation has been applied to hundreds of physical stimuli such as light (brightness), sound (volume), heat, cold, and pressure. One of the early findings in psychophysics was that the perception of physical stimuli is regular, but that the relationship between the perceived strength and physical strength of most stimuli is non-linear, or more precisely, exponential (Stevens 1957).

It is easiest to illustrate this discovery graphically. Given that the relationship of interest is between the physical strength of the stimulus and the perceived strength,

one can plot the physical strength on the x-axis and the perceived (estimated) strength on the y-axis. If the relationship between physical and perceived strength were linear (as it is in the line length example above), the graph would be a straight line.[1]



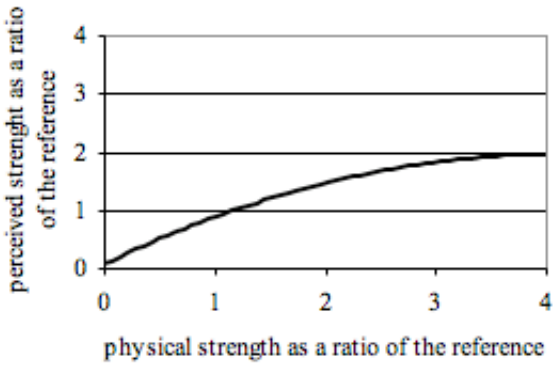
Making a similar graph of an exponential relationship results in a parabola, whose orientation is determined by the exponent. For instance, an exponent of 2 yields the following graph with an upward opening parabola:



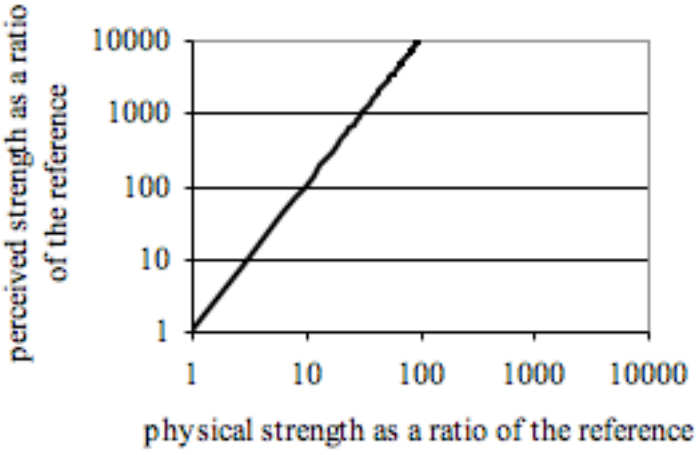
[1]This abstracts away from the fact that the slope of this line could be a value other than 1. A slope other than 1 would indicate a regular (lawful) distortion of the perception.

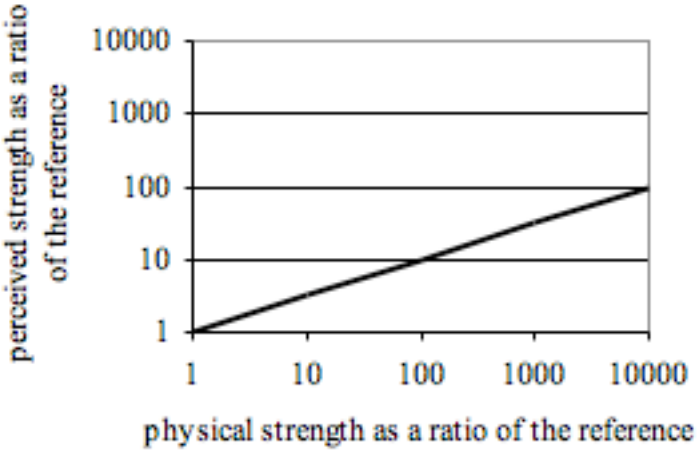
The graph can be read in the following way: a stimulus that is physically identical to the reference (a value of 1 along the x-axis) will be perceived as being equally as strong as the reference (a value of 1 along the y-axis). However, a stimulus that is physically twice as strong as the reference (a value of 2 along the x-axis) will be perceived as 4 times as strong as the reference (a value of 4 along the y-axis). In general, exponential relationships in which the relationship is greater than 1 have two consequences: (i) physical stimuli larger than the reference will be overestimated, while physical stimuli smaller than the reference will be underestimated, and (ii) the larger the difference between the stimulus and the reference, the larger the overestimation (or underestimation if the stimulus is smaller than the reference).

For exponential relationships characterized by exponents less than 1 the opposite is true: (i) physical stimuli larger than the reference will be underestimated while physical stimuli smaller than the reference will be overestimated, and (ii) the larger the difference between the reference and the stimulus, the larger the underestimation (or overestimation if the stimulus is smaller than the reference). This can be illustrated with the following graph of an exponential relationship of .5 (rightward opening):



For clarity, exponential relationships like the ones above can be graphed on a logarithmic scale. The resulting graph is a straight line with a slope equal to the exponent of the relationship. In the first example the exponent was 2, therefore the same relationship graphed on a logarithmic scale would result in a straight line with a slope of 2. In the second example the exponent was .5, therefore the same graph on a logarithmic graph would result in a straight line with a slope of .5:





The fact that the relationship between physical stimuli and perception is exponential is known as the Psychophysical Power Law (Stevens 1957). Each physical

stimulus has its own characteristic exponent; some of these exponents are provided in the following table, adapted from Lodge 1981, along with the ratios of perception when the stimulus is 2, 3, and 4 times the strength of the reference:

Table 2.1. Characteristic exponents for various physical stimuli

Stimulus	Exponent	2x	3x	4x
brightness	0.5	1.4	1.7	2.9
volume	.67	1.6	2.1	2.5
vibration	.95	1.9	2.8	3.7
length	1.0	2	3	4
cold	1.0	2	3	4
duration	1.1	2.1	3.3	4.6
finger span	1.3	2.5	4.2	6
sweet	1.3	2.5	4.2	6
salty	1.4	2.6	4.7	7
hot	1.6	3	5.8	9.2

2.1.2 Magnitude estimation in linguistics

The extension of magnitude estimation to linguistics is superficially straightforward. Participants are presented with a pair of sentences. The first is the reference sentence, and has a value associated with its acceptability (in this example, 100). The acceptability of the second sentence can then be estimated using the acceptability of the first. If the sentence is two times more acceptable than the reference, it would receive a value twice the value of the reference (e.g., 200). If the sentence is only half as acceptable, it would receive a value half that of the reference (e.g., 50):

Reference: What do you wonder whether Mary bought?
Acceptability: 100

Item: What did Lisa meet the man that bought?
Acceptability: _____

Again, by using the same reference sentence to judge several other sentence types, the relative acceptability of those sentence types can be estimated in units equal to the acceptability of the reference sentence.

The obvious question, then, is why linguists should adopt magnitude estimation over other more familiar measurement techniques, such as yes/no tasks or 5- (or 7-) point scale tasks. There are two major purported benefits of magnitude estimation for linguistics. First, the freedom of choice offered by the real number line means that respondents can report any difference between stimuli that they find relevant. This is in contrast with categorization tasks such as the yes/no or scale tasks in which the number of choices is set by the experimenter. Categorization tasks increase the likelihood of losing relevant data by forcing respondents to group potentially different stimuli into the same category (e.g., the alphabetic grading system in the U.S.). Magnitude estimation sets no such categories, and the theoretical infinity of the real number line ensures that respondents always have the option of creating an additional distinction. While the question of whether scale tasks actually lose linguistically relevant data is an empirical one (see Schütze 1996 for a discussion of the effect of different scale sizes), there is growing evidence that the answer may be yes: Bard et al. 1996 demonstrate that respondents tend to distinguish more than 7 levels of acceptability when given a magnitude estimation task, and Featherston

2005b demonstrates that magnitude estimation reveals distinctions among sentences that are classified as 'good' in a yes/no task.

The second purported benefit of magnitude estimation concerns the nature of the data itself: it has been claimed that magnitude estimation data can be analyzed using parametric statistics, while yes/no and scale studies are only amenable to non-parametric statistics (Bard et al. 1996, Keller 2000). Parametric statistics are so named because they assume the data has the following 5 properties:

1. The sample is chosen at random
2. The observations are independent of each other
3. The variances of each group are equal
4. The intervals between data are meaningful
5. The observations are normally distributed

The first three assumptions are independent of the task, and generally under the control of the experimenter, therefore not much more will be said about them.[2] As

[2]However, that is not to say that they are necessarily satisfied by acceptability studies. For one, the assumption of random sampling is hardly ever satisfied: most participants are college students who have actively sought out the experimenter. The assumption of independence of observations is of considerable concern, especially given the claims tested in chapter 4 that the repetition of structures affects acceptability. As for the assumption of equal variances, no large scale studies of the variance of acceptability across structures has ever been conducted, although there is no a priori reason to believe that variance should be affected by structure type.

for the final two assumptions, the claim for magnitude estimation has been that it satisfies them while standard scale tasks do not.

In general there are four types of scales in statistics: nominal, ordinal, interval, and ratio. Nominal scales assign categorical differences to data, but do not specify any ranking, ordering, or distance among the categories. The visible color names are a good example of a nominal scale: while there is a physical spectrum of color, the names themselves do not specify this order; the names simply categorize chunks of the spectrum into groups, with no regard for their physical relationship (hence the need for mnemonics such as ROY G BIV). The yes/no task of acceptability uses a nominal scale. Ordinal scales add ordering to nominal scales, but still do not specify the distance between the categories. Again, the alphabetic grading system in the U.S. is a good example: there is a definite ordering of the grades, but the distance between them is not necessarily stable, as anyone who has dealt with multiple graders or multiple sections of a class can attest. The scale tasks of linguistic acceptability are ordinal scales: they specify the order of structures, but it is not clear that the differences between points on the scale are stable either within or across participants.

Interval scales, which are assumed by parametric statistics, specify both the ordering of items and the distance between them, but not the location of 0. The temperature scales Fahrenheit and Celsius are examples of interval scales: the distances between any two degree marks on a thermometer are identical, so the intervals are stable, but there is no meaningful 0 point. The 0 point on the Celsius scale is arbitrarily set as the point at which water freezes, whereas the 0 point on the Fahrenheit scale is even more arbitrarily set as 32 degrees below the point at which water freezes.

Because the 0 points are arbitrary, it is not the case that 64 degrees F is twice as much as 32 degrees F; all we can say is that 64 degrees F is 32 degrees higher than 32 degrees F, because 0 degrees F is not the absence of temperature. Magnitude estimation scales of acceptability have been claimed to be interval scales: the distance between any two structures is measured in units equal to the acceptability of the reference, therefore the intervals are stable. It is not clear what it means to say that there is 0 acceptability, so the 0 point of acceptability must be arbitrary.

Ratio scales are also amenable to parametric statistics because they also assume stable intervals, but add a meaningful 0 point. The Kelvin temperature scale is a good example of a ratio scale: 0 Kelvin is commonly called 'absolute zero' because it represents the absence of all temperature. On this scale 100 Kelvin is twice as much as 50 Kelvin because 0 Kelvin is meaningful. The Kelvin scale confirms that 64 degrees F is not twice as much as 32 degrees F: 32 degrees F is about 273 K, and 64 F is about 291 K. The magnitude estimates from psychophysics are all on ratio scales as there is a meaningful 0 point for measures of brightness, volume, pressure, et cetera.

In addition to providing interval level data, it has also been suggested that magnitude estimation provides normally distributed data (as opposed to the ordinal data from scale tasks, which is never normally distributed). The normal distribution (or Gaussian distribution, also commonly known as the bell curve) is a crucial assumption underlying parametric hypothesis testing. At its heart, hypothesis testing asks the following question: given these (two) samples (or sets of data), how likely is it that they come from the same population? In behavioral studies, it is generally

assumed that if the likelihood is less than 5% that the two samples come from the same population, then it can be concluded that they do not (hence, a significant difference).

While the mathematics of each statistical test is different, the logic is the same. Each sample is an approximation of the population. If we can estimate the population mean from the sample mean, we can then compare the population mean estimates from each sample mean to determine if they are from the same or different populations. Enter the normal distribution: because statisticians have determined the properties of the normal distribution, if a sample is normally distributed, the properties of the normal distribution can be used to estimate the population mean from a sample mean. So in a very real sense, the numbers that are used to determine statistical significance (the estimated population means for each sample) are contingent upon the samples being normally distributed. If the sample is not normally distributed, the estimated population means will be incorrect, and the likelihood that they are from the same sample will be incorrect. Therefore, the fact that magnitude estimation has been purported to provide normally distributed data is an important innovation for studies of acceptability.

In the following sections, these two claims about magnitude estimation data will be investigated through a meta-analysis of the experiments that will presented throughout the rest of the dissertation:

1. Magnitude estimation yields interval level data

2. Magnitude estimation yields normally distributed data

Unfortunately, the results suggest that the data, although internally consistent, is not interval level at all, and while the data is normally distributed, the standard analysis techniques applied in the syntax literature destroy the normality prior to statistical analysis. These facts will be considered with respect to the two major goals of syntactic theory presented in the first chapter to determine whether we need to reconsider any of the conclusions that have been drawn from magnitude estimation data. One note is in order before proceeding: Because of the quantity of data being meta-analyzed (a necessity given the goal) and because the details of these experiments will be presented in subsequent chapters, these sections will provide only a minimum amount of experimental detail. The reader is referred to the subsequent chapters for more detailed experimental reporting.

2.2 The interval scale of linguistic magnitude estimation

2.2.1 Internal consistency and cross-modality matching

One of the most obvious differences between psychophysical and linguistic magnitude estimation concerns the physical measurement of the stimulus: in psychophysics, the stimulus can be objectively measured, independently of the observer; acceptability cannot be measured independently of the observer. Together with the claim that magnitude estimation of acceptability yields interval level data, this raises the obvious question: with no possibility of external validation, how can we be sure that participants are applying consistent intervals?

Bard et al. 1996 address this concern directly. They demonstrate that individual participants are internally consistent in their judgments with a methodology borrowed from psychophysics called cross-modality matching. The logic of cross-modality matching is straightforward: participants are presented with a set of stimuli and asked to estimate their magnitude using one modality, for instance real numbers as presented above. Then, the same participants are asked to estimate the magnitude of the same stimuli using a different modality, such as drawing lengths of lines. For this second modality, rather than assigning numbers with the correct proportions, they would be asked to draw lines with the correct proportions. The two modalities can then be compared: if a participant is internally consistent, there will be a direct relationship between the responses in each modality. In concrete terms, imagine a reference sentence whose acceptability is set using the number modality as 100. A stimulus sentence that is two times as acceptable would be assigned a value of 200. Then imagine the same reference sentence and stimulus sentence pair, in the line length modality. The reference sentence's acceptability could be assigned a line length of 100 points (about 1.4 inches). If the participant is being internally consistent, they should then assign a line length of 200 points (about 2.8 inches) to the stimulus sentence, or in other words, there should be a direct relationship between the two modalities. Bard et al. 1996 find exactly that internal consistency.

2.2.2 Interval regularity and the effect of the reference sentence

Of course, internal consistency within participants does not guarantee that the intervals in magnitude estimation are regular, just that the non-regularity is consistent: scale tasks are known to result in non-regular intervals while still yielding consistent results within participants. To test the regularity of intervals, two experiments will be compared. The experiments are identical in nearly every way: they test the same conditions, use the same items, are presented in the same order, et cetera. The only difference between the two experiments is that they each use a different reference sentence. The first experiment uses an If-Island violation as the reference and the second uses a Coordinate Structure Constraint (CSC) violation as the reference.[3] Crucially, the Coordinate Structure Constraint is one of the conditions in each experiment. Therefore, the acceptability values for each of the conditions in the first experiment can be translated into values of the Coordinate Structure Constraint - in essence, predicting the values that should be obtained in the second experiment if the intervals are indeed regular.

Table 2.2. Reference sentences for each experiment

If Island	What did you ask if Larry had bought?
CSC violation	What did you say Larry bought a shirt and?

Participants were presented with 5 tokens of each violation type, along with 10 grammatical fillers, for a total of 50 items. Mean values for each condition were obtained by first dividing the response by the reference value (to normalize the scale

[3]The If-reference was chosen because it is a frequently used reference in the literature (e.g., Keller 2000).

Table 2.3. Conditions in both experiments

Adjunct Island	What does Jeff do the housework because Cindy injured?
CSC violation	What did Sarah claim she wrote the article and ?
Infin. Sent. Subject	What will to admit in public be easier someday?
Left Branch Condition	How much did Mary saw that you earned money?
Relative Clause	What did Sarah meet the mechanic who fixed quickly?
Sentential Subject	What does that you bought anger the other students?
Complex NP Constraint	What did you doubt the claim that Jesse invented?
Whether Island	What do you wonder whether Sharon spilled by accident?

to ratios of the reference sentence), then determining the mean of each condition for each participant. The grand mean of the participant means was then calculated for each condition:

Table 2.4. Conditions means for If-reference

Condition	Mean
Whether Island	0.89
Complex NP Constraint	0.79
Left Branch Condition	0.70
Relative Clause	0.70
Adjunct Island	0.69
CSC violation	0.65
Sentential Subject	0.58
Infin. Sent. Subject	0.52

As the table indicates, the mean for the CSC condition in the If-reference experiment was 0.65 times the reference. This value can then be set as 1 to calculate the predicted values for the other conditions in an experiment in which the CSC is the reference (i.e., has a value of 1):

Table 2.5. Predicted means for CSC-reference based on If-reference results

Condition	Predicted Mean
Whether Island	1.37
Complex NP Constraint	1.21
Left Branch Condition	1.08
Relative Clause	1.08
Adjunct Island	1.07
CSC violation	1.00
Sentential Subject	0.89
Infin. Sent. Subject	0.80

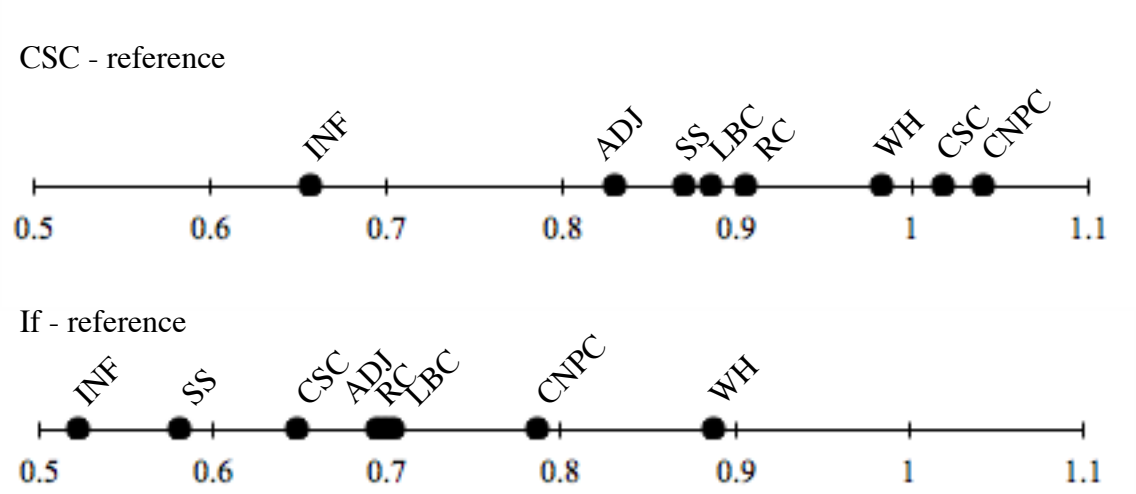
This can then be compared to the experimentally obtained values:

Table 2.6. Predicted and actual means for CSC-reference

Condition	Predicted	Actual
Whether Island	1.37	0.98
Complex NP Constraint	1.21	1.04
Left Branch Condition	1.08	0.89
Relative Clause	1.08	0.90
Adjunct Island	1.07	0.83
CSC violation	1.00	1.02
Sentential Subject	0.89	0.87
Infin. Sent. Subject	0.80	0.66

As the table indicates, the predictions do not hold experimentally. In fact, the relative order of the absolute values of the means in the CSC experiment are different from the relative order of the absolute values of the means in the If experiment:

Figure 2.1. Relative means for both experiments



These results indicate that the intervals in the first experiment are different from the intervals in the second experiment. This suggests that either (i) magnitude estimation of acceptability is not interval level data, or (ii) magnitude estimation of acceptability is not real magnitude estimation (i.e., the magnitudes are not estimated using the reference sentence as a unit). Either way, this suggests that the intervals in magnitude estimation of acceptability are not regular.[4]

2.2.3 Interval irregularity and its effect on theoretical syntax

The irregularity of the intervals in the previous subsection is not that surprising: it has already been admitted that there is no meaningful 0 point for acceptability,

[4]In fact, at least one published study has implicitly accepted this as true. Featherston 2005a normalized participants' responses using z-scores, a method of comparing participants' responses based on how far each response is from that particular participant's mean. In Featherston's words, "this effectively unifies the different scales that the individual subjects adopted for themselves."

yet despite this fact, the magnitude estimation task explicitly asks participants to give ratio judgments. Bard et al. 1996 rationalize this by suggesting that the ratio scale created by participants assumes an interval scale, therefore the data will be interval level:

> ... providing that subjects' abilities are as great as we have supposed, attempt to say which sentence is 1.5 times as acceptable as another, and which .6 times as acceptable, and so forth, can at least give us the interval scales that we need.

However, it is not clear how sound this logic is. If there is no true 0 point, but participants are still asked to respond as if there were one, are we not asking the participants to estimate a 0 point? Even in the unlikely event that all participants estimate the same 0 point for a given reference sentence, it is still possible that they will estimate a different 0 point for a different reference sentence, and that the two 0 points might be irregularly related, resulting in meaningless intervals outside of a given experiment. Such a situation is not that unlike the scale tasks that magnitude estimation was meant to replace.

While the lack of interval level data may be distressing to the statistically conscientious, the real question for the practicing linguist is whether this fact affects the interpretation of magnitude estimation data with respect to the two driving questions of theoretical syntax:

1. What phenomena count as grammatical knowledge?

2. What form does grammatical knowledge take?

Ultimately, the answer to both of these questions will be empirically determined as the body of magnitude estimation results grows. These two experiments can only make a first suggestion as to what the answers may look like.

As to the first question of whether this irregularity will affect analyses of individual phenomena, the answer appears to be yes and no. Any analysis that is predicated upon the absolute values of the measurements in magnitude estimation will obviously be affected. The absolute values of differences between conditions is strongly dependent upon the choice of reference sentence, and at least from these two studies, there does not appear to be a regular relationship between the values of different reference sentences. For the most part this will have little impact on theoretical syntax, as the field in general has never attempted to make numerical predictions regarding the acceptability of sentences. There have been some recent attempts to calculate weightings for different syntactic constraints (e.g., Keller 2003), which may require further testing using a variety of reference sentences to determine a range for the weightings.

For analyses that are predicated upon relative acceptability, as the majority of theoretical studies are, these experiments suggest little reason for concern. The statistically significant relative comparisons among these islands are maintained across these two experiments (except for comparisons involving the CSC, which will be addressed shortly). This can be seen both in graphs of the relative acceptability of the conditions in both experiments, which show the same general shape, and in the significance of the statistical comparisons, which are unchanged:

Figure 2.2. Pattern of acceptability for both ratios and logs

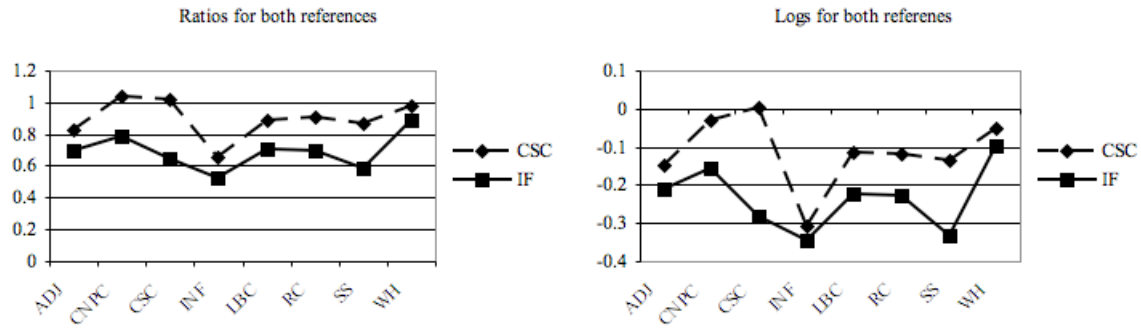


Table 2.7. Paired samples t-test results for both experiments

	If reference				CSC reference			
	ratios		logs		ratios		logs	
Contrast	t	p	t	p	t	p	t	p
Adj - Inf	3.86	.001	3.80	.001	4.50	.001	4.50	.001
CNPC - SS	3.58	.001	3.95	.001	3.1	.002	3.14	.002
WH - RC	4.88	.001	5.30	.001	1.70	.05	2.09	.023

As to the second question of whether the interval irregularity adds to our knowledge of the form of grammatical knowledge, it is interesting to note an unexpected difference between the two experiments. In the If-reference experiment, the CSC is one of the least acceptable violations; indeed, only the two Sentential Subject Islands are numerically less acceptable. However, in the CSC-reference experiment, all of the violations except the CNPC Island are judged less acceptable than the CSC-reference. In a certain sense, it seems as though the reference sentence was judged as more acceptable than the other violations by virtue of being the reference sentence, as if its inherent acceptability were ignored. This is also true of the If-reference experiment: all of the violations were judged as worse than the If-reference sentence. Of course, in the case of the If-reference, we have no independent evidence that it

should be judged worse than some of the other violations. However, in the case of the CSC, the first experiment suggests that we should expect it to be near the bottom of the acceptability hierarchy, contrary to the observed results.

One claim about the nature of grammaticality that has been put forward is that the continuous measurement (sometimes called gradience) in judgments that comes from magnitude estimation is indicative of the gradience in grammatical knowledge (e.g., Keller 2000, Sorace and Keller 2005). Of course, the fact that the results from magnitude estimation are continuous is unsurprising given that the task is a continuous task, much the way it is unsurprising that yes/no tasks provide evidence for binary grammatical knowledge. However, it has been claimed that one of the virtues of magnitude estimation tasks is that they do not impose any conception of 'grammatical' or 'ungrammatical' on the participant. As Featherston 2005a said:

> Subjects are asked to provide purely comparative judgments: these are relative both to a reference item and the individual subject's own previous judgments, but at no point is an absolute criterion of grammaticality applied.

However, in these experiments, we have the suggestion of a grammatical-ungrammatical distinction being brought to the task *by the participants*: these results suggest that ungrammatical sentences are judged as equal to or less acceptable than the reference sentence regardless of its relative acceptability to the other violations. In fact (and unsurprisingly), the grammatical fillers in these experiments are judged as more acceptable than the reference as well. Only further studies can demonstrate

whether this is in fact a stable trend across all possible reference sentences or just a quirk inherent to the two under consideration here. For now, though, this fact suggests that the despite the lack of explicit grammaticality in the magnitude estimation task, participants' responses are affected by the binary grammaticality of the sentences being investigated - a potential piece of evidence for a binary form of grammatical knowledge.

2.3 The distribution of acceptability judgments

The second statistical assumption that magnitude estimation data is purported to satisfy is that the distribution of responses be normal. Recall that in psychophysical magnitude estimation, empirical observation demonstrated that the perception of physical stimuli was not normal. In fact, the distribution of psychophysical magnitude estimates was *log-normal*, in that it was characterized by a power law. The broad question addressed by this section, then, is whether similar empirical observation will demonstrate that acceptability magnitude estimates are normal (as assumed by parametric statistics), or perhaps even log-normal (as was found in psychophysics).

2.3.1 The log transformation

Before addressing the broad question about the nature of the distribution of acceptability magnitude estimates, it is necessary to discuss the commonly accepted data analysis methods of linguistic magnitude estimation. The issue at hand is that most of the magnitude estimation experiments in the syntactic literature do not report

the raw ratio responses given by the participants; instead, these experiments report the (natural) logs of the ratios. Transforming the ratios into logs (a log transformation) significantly affects the distribution of the responses; therefore, it is necessary to deconstruct the logic behind this transformation before addressing the underlying issue of the distribution of acceptability magnitude estimates.

There appear to be several reasons given in the literature for why the log-transformation is applied to magnitude estimates prior to analysis. For instance, Bard et al. 1996 offer two reasons:

> Logs are used both to keep the scale manageable in the presence of very large numbers some subjects used and also to provide a straightforward way of dealing with the judgments of proportions: when exponentiated, the difference between log estimates provides the ratio of the acceptability of the two versions of the sentence.

The first reason, to keep large numbers manageable, is a non-technical reference to the fact that magnitude estimation data given in terms of numbers will by definition have a rightward skew: the number line used by participants is infinite in the rightward (increasing) direction, but bounded by 0 to the left. The log transformation limits the effect of large numbers, therefore keeping the data 'manageable' by limiting the effect of outliers. However, there are many other methods for dealing with outliers. In fact, the log transformation is a rather extreme measure for simple outlier removal: as the name implies, it changes the entire nature of the distribution. Weaker outlier removal techniques can achieve the same goal without affecting the overall distribution

of the data. The second reason, because exponentiating logs yield ratios, is true, but is peculiar to the goals of the study being analyzed: the reported study focused on difference scores between two sentences. Exponentiating difference scores does indeed reveal the ratio of the two original scores, which in this case saves one step of the analysis. However, reporting the original ratios of each condition would just as easily make both the differences and the ratio relationships apparent (with one simple calculation each), without recourse to the log transformation.

Keller 2000 offers two additional reasons for applying the log transformation to acceptability magnitude estimates:

> This transformation ensures that the judgments are normally distributed and is standard practice for magnitude estimation data (Bard et al. 1996, Lodge 1981).

The first reason is exactly the empirical question that needs to be answered. If it the case that acceptability judgments are log-normal, the log transformation will be necessary to create a normal distribution of the data. The second reason, that it is standard practice in magnitude estimation, is a more complicated issue. As we have just seen, the reasons offered by Bard et al. 1996 may not be the most compelling. The other referenced work, Lodge 1981, is a brief how-to guide for applying magnitude estimation to questionnaires in sociology. The methodology in that guide is identical to the methodology in psychophysics. Although there is a log transformation in the psychophysical/sociological data analysis, it does not have the same effect as it does in linguistic magnitude estimation. A comparison of the two methodologies will make

this more apparent.

Recall that psychophysical studies were interested in the ratios provided by participants. The fact that there is a 0 point in physical perception meant that these ratios were meaningful, and the fact that the physical stimuli could be externally measured meant that the perceived ratios could be compared to the actual ratios. Because ratios are a form of multiplication, the appropriate *central tendency* for ratios is not the more familiar arithmetic mean (commonly called the *average*), but rather the geometric mean.[5] Now, there is a simple algorithm for determining the geometric mean, and it involves the log transformation:

1. Log transform the values.
2. Calculate the arithmetic mean of the logs.
3. Exponentiate (raise 10 to the power of the number) the arithmetic mean.
4. The result of the exponentiation is the geometric mean.

As such, in order to determine the geometric mean of magnitude estimates, psychophysicists did apply log transformations, but these transformation were then 'undone' by the exponentiation. So while the magnitude estimates were indeed log-normal (as proven by the Power Law), the log transformation was not used to normalize the distribution - after exponentiation, the distribution was again log-normal

[5]If the arithmetic mean answers the question *What is the value I need to add to itself X times to achieve this sum?*, the geometric mean answers the question *What is the value I need to multiply by itself X times to achieve this product?*. In everyday life, the geometric mean is the appropriate way of figuring out an average rate of return on an investment.

(as demonstrated by the use of logarithmic graphs).

In linguistic magnitude estimation, a similar procedure is followed: the participants' responses are log transformed, and the arithmetic mean of the logs is calculated. However, the parallels with the psychophysical analysis end there. The logs are used in the subsequent statistical analysis, not the geometric means that can be calculated from the logs. The question, then, is why linguistic magnitude estimation uses the log transformation if it is not to calculate geometric means of the ratios, with a natural follow-up question being why linguistic magnitude estimation does not calculate the geometric means of the ratios.

We have already discussed one possible reason for not analyzing the geometric means of the ratios: acceptability ratios have no meaningful 0 point, so the ratios themselves are not meaningful. However, if we admit the ratios are meaningless, then we must ask why it is that the task *asks participants to determine ratios*. Log transforming the meaningless ratios will not make them more meaningful. As for the logic behind the log transformation, we are left with Keller's reason: the acceptability judgments are log-normal to begin with, so log transforming them is necessary to make them normal.

2.3.2 The distribution of judgments in this dissertation

This subsection directly addresses the question of whether acceptability magnitude estimates are normally distributed, log-normally distributed, or neither, by reporting normality tests (specifically the Shapiro-Wilk test) for both the raw (ra-

tio) and log transformed responses to four of the experiments reported in subsequent chapters. As we shall see shortly, the tests reveal that the raw magnitude estimates are normally distributed when multiple responses are collected for each participant, but that the log-transformation destroys this normality. This suggests that the log-transformation is not only unnecessary from a normality perspective, but also inappropriate as a method for removing outliers as suggested by Bard et al. 1996. More conservative techniques such as trimming or Winsorization may be more appropriate.

2.3.2.1 Multiple observations per participant

The first two experiments to be analyzed are the two Island studies reported in the previous section. In these studies, participants provided 5 observations per condition, which were averaged for each participant to reduce the influence of outliers. Each experiment consisted of 8 ungrammatical conditions and 1 grammatical condition, for a total of 18 conditions across the two experiments. The Shapiro-Wilk test was applied to each condition for both ratios and logs. The results of the normality tests for both types of data (ratio and log) can be summarized in the following table:[6]

Table 2.8. Qualitative normality results for multiple observations per participant

	ratios	logs
normal	16	4
non-normal	2	14

[6]Due to the sheer number of Shapiro-Wilk tests performed, the results will only be reported with qualitative summaries. The statistics for each test are reported at the end of the chapter.

As the table suggests, the raw magnitude estimates are overwhelmingly normal. In fact, the two non-normal conditions are easily explained. The first non-normal condition is the CSC condition within the CSC-reference experiment. The fact that the responses to this condition are non-normal is unsurprising given that it is structurally identical to the reference sentence. The second non-normal condition, the grammatical condition in the CSC-reference experiment, actually becomes normal with the elimination of 2 outliers, suggesting that it is only superficially non-normal. Perhaps more interesting that these two exceptions to the overwhelming normality of the ratio judgments is the fact that the logs of the ratios are overwhelmingly non-normal. These two facts combined suggest, at least for these two experiments, that the log transformation is unnecessary (and inappropriate) as the ratio judgments are already normally distributed.

2.3.2.2 Few observations per participant, large sample

The second pair of experiments differ from the first pair in that multiple observations were not collected for each participant. However, these two experiments did involve a large number of participants (86 and 92), resulting in an equivalent number of total observations. While the lack of multiple observations per participant means that outliers have a greater effect on normality, the unusually large number of total observations ensures that once the outliers are accounted for, the distribution of the resulting scores will not be non-normal due to insufficient measurements. The two experiments focused on 12 ungrammatical structures and 6 grammatical structures,

for a total of 18 conditions, as summarized in the following table:

Table 2.9. Qualitative normality results for few observations per participant, large sample size

	ratios	logs
normal	1	0
non-normal	17	18

Unlike the first pair of experiments, the raw ratios of these two experiments are overwhelmingly non-normal. Interestingly, the logs are also overwhelmingly non-normal. This suggests that the non-normality in the ratios is not because they are log-normal. Given the lack of multiple observations per condition per participant, the non-normality may be due the influence of outliers. Whatever the ultimate source of the non-normality of the raw ratios, together with the results from the first pair of experiments, these results strongly suggest that the log transformation is inappropriate in linguistic magnitude estimation: there is no evidence that acceptability magnitude estimates are log-normal, and in fact, there is evidence that the untransformed ratios tend to be normal when the influence of outliers is minimized.

2.3.2.3 Few observations per participant, small sample

The final 4 experiments to be considered differ from the previous experiments along both dimensions: only 1 observation per participant and a relatively small number of participants (between 20 and 26). The relatively low number of total observations means that these experiments are more likely to be influenced by outliers, and therefore less likely to be normal. However, we can still compare the distribution of the raw ratios to the distribution of the logs to determine the validity of the log

transformation:

Table 2.10. Qualitative normality results for few observations per participant, small sample size

	ratios	logs
normal	19	35
non-normal	43	27

For the raw ratios, there are significantly more non-normal conditions than normal conditions by Sign Test ($p<.003$), which is unsurprising given the increased influence of outliers in these designs. And while the log transformation does increase the number of normal conditions from 19 to 35, there still are not significantly more normal conditions than non-normal conditions by Sign test ($p<.374$). So while the log transformation does improve the overall normality in these designs, there is no evidence to suggest that the general distribution of acceptability magnitude estimates are log-normal.

2.3.2.4 The overall distribution of judgments

The results from analyzing the distribution of responses in these experiments is summarized in the following table, including the results of a Sign Test:

Table 2.11. Qualitative normality results for multiple observations per participant

	ratios			logs		
	normal	non-normal	p	normal	non-normal	p
many observations	16	2	.001	4	14	.031
few observations, large N	1	17	.001	0	18	.001
few observations, small N	19	43	.003	35	27	.374

The general picture that emerges is as follows:

1. The raw ratios are normally distributed in designs that minimize the influence of outliers.

2. In designs that do not minimize the influence of outliers, the raw ratios are non-normal, but so are the logs.

3. This suggests that the non-normality is more likely to be due to outliers than log-normality of acceptability magnitude estimates.

Because the distribution of acceptability magnitude estimates is not log-normal, and in fact tends to be normal in designs that minimize the influence of outliers, it is more appropriate to apply non-transformation outlier correction methods such as trimming or Winsorization than the much stronger log transformation.

2.3.3 The effect of the log transformation on theoretical syntax

The results of the previous subsection seem to suggest that the log transformation, although standard in the linguistic magnitude estimation literature, is not an appropriate transformation for acceptability magnitude estimates. In fact, its stated purpose, to ensure the normality of the results, seems unnecessary given the normality of the responses prior to the transformation. The question, then, is whether the application of the log transformation has affected the results of magnitude estimation experiments in a way that affects syntactic theory; or in other words, does the log transformation affect the first major question driving theoretic syntax, what phenomena constitute grammatical knowledge?

In general, F-tests such as ANOVA are robust to violations of normality when the the non-normal distributions in question are identical (i.e., non-normal in the same way) (Wilcox 1997). In the case of acceptability magnitude estimates, it is true that the log transformation creates a non-normal distribution, but it is also true that the log transformation is applied to all conditions, making it likely that the non-normal distributions are similarly non-normal. Given the robustness of F-tests to identical non-normality, we would not expect the log transformation to significantly alter the results of these tests. This was tested using the experiments from this dissertation by comparing the raw ratios and logs: while absolute p-value and effect sizes did change after the log transformation, the qualitative results (the presence or absence of a significant effect) did not change.

2.4 The real benefits of magnitude estimation

One of the major claims about magnitude estimation is that it provides data that is more amenable to parametric statistics than other standard judgment collection techniques, such as scale tasks. However, the results of this chapter suggest that the quality of data from magnitude estimation suffers from some of the same drawbacks as data from scale techniques. For one, it has been claimed that magnitude estimation yields interval level data, as is required for parametric statistics. Yet a comparison of two identical magnitude estimation experiments suggests that the intervals constructed by participants are not regular. While this result is not entirely surprising given that participants are actually asked to produce ratio responses to a

stimulus with no meaningful 0 point, it does suggest that magnitude estimation data is ordinal, just like the data from scale tasks. In fact, the distribution of responses with respect to the reference sentence suggests that participants are not performing magnitude estimation at all, but rather a standard scaling technique in which grammatical sentences are placed above the reference, and ungrammatical below the reference. While this is interesting evidence of a psychological difference between grammatical and ungrammatical sentences, it is also very similar to standard scale tasks in which the mid-point of the scale marks the difference between the two categories.

It has also been claimed that the continuous measures produced by magnitude estimation are normally distributed following the 'standard' log transformation. Yet normality tests on several experiments revealed that magnitude estimation responses are already normal *before the log transformation* in designs that minimize the influence of outliers. And although the untransformed responses are non-normal in designs that do not minimize the influence of outliers, the log transformation does not make these distributions normal, suggesting that simple outlier removal would be a more appropriate procedure. Taken together, these findings suggest that the absolute numbers associated with magnitude estimation are unreliable: the lack of regular intervals indicates that the magnitude of differences is unreliable, and the lack of normality after the log transformation suggests that the exact statistics may be unreliable. However, comparing statistical analyses on both raw and log transformed data, and on experiments with different reference sentences (hence different intervals), has demonstrated that the qualitative effects are not affected (i.e., statistically significant results remain significant and no new significant results emerge).

These results suggest that in many ways magnitude estimation of acceptability is similar to standard 5- or 7-point scale rating tasks of acceptability. In both paradigms, the relative differences between structures is reliably determined using standard parametric statistical tests such as the ANOVA, but this has more to do with the robust nature of the effects (and the statistical tests) than the precision of the measuring instrument. In the end, these results suggest that the real benefit of magnitude estimation rests not in the data it yields, but in the freedom it gives participants. Bard et al. 1996 and Featherston 2005b have already demonstrated that there are significant differences in acceptability that appear reliably with magnitude estimation but nevertheless may not be apparent to other rating techniques. Furthermore, as Featherston 2005a points out, there is no explicit or implicit distinction between grammatical and ungrammatical sentences in the magnitude estimation task itself: every sentence is treated equally as a receiving an estimated acceptability based on the reference sentence. Consequently, the apparent distinction between grammatical and ungrammatical sentences discussed above must have been introduced by the participants themselves. If this finding is found to hold for a variety of reference sentences, then magnitude estimation may become new psychological evidence for the categorical distinction between grammatical and ungrammatical sentences.

The final question, then, is what this means for the practice of acceptability judgment collection and its impact on our understanding of grammatical knowledge. The bottom line appears to be that the absolute values of judgments, and the gradience implied by them, are not necessarily reliable. The relative differences are reliable,

despite the inappropriate application of the log transformation.[7] So as long as effects are defined relative to control conditions, there is no cause for concern. There is reason to be cautious in drawing conclusions of gradience from these results (after all, gradience is built into the task), especially if precise values or weights are assigned to that gradience, as the values may change with the reference sentence. However, the emergence of a categorical grammaticality distinction on the relative judgments may ultimately provide a valuable new insight into the nature of grammatical knowledge.

[7]Even though the log transformation is not ultimately causing any statistical harm, given that it obscures the intentions of the participants (a sentence judged as twice as acceptable as the reference becomes .3 after the log transformation), it would probably be worth abandoning the log transformation in favor of the geometric means. However, for consistency with the standard practices of the field, the results for the rest of the dissertation will still include the log transformation.

Chapter 3

Context, acceptability, and grammaticality

One of the recurrent questions facing syntactic research is whether the fact that acceptability judgments are given for isolated sentences out of their natural linguistic context affects the results. This chapter lays out several possible ways in which the presence or absence of context may affect the results of acceptability studies, and what effect each could have on theories of grammatical knowledge. Because the effect of context can only be determined empirically for individual grammatical principles on a case by case basis, this chapter then presents 3 case studies of the effect of various types of context on properties of wh-movement in an attempt to determine i) if there is an effect, and ii) if so, what the underlying source of the effect may be. These case studies serve a dual purpose: they serve as a first attempt at systematically investigating the effect of context on properties of wh-movement, and, given that wh-movement serves as the empirical object of study in the rest of the dissertation, they serve as a set of control experiments to determine whether context should be included as a factor in experiments in subsequent chapters. As will soon become apparent, these experiments suggest that at least the types of context studied here do not affect major properties of wh-movement such as Island effects and D-linking effects, indicating that context need not be included as a factor in subsequent studies of Island effects in the rest of the dissertation.

3.1 The complexities of context

As Schütze 1996 points out, the acceptability of a sentence is dependent upon finding a context in which its meaning is appropriate. In standard acceptability judgment tasks, especially informal experiments conducted by syntacticians daily, overt linguistic context is rarely supplied, except perhaps in order to distinguish the meanings associated with ambiguous structures. Given that acceptability as a property is dependent on context, and given that context is rarely supplied during the collection of acceptability judgments, it is no wonder that both linguists and psychologists have questioned the reliability of acceptability data:

> Although grammaticality judgments are considered an extremely rich source of data, it has long been evident that introspections about decontextualized, constructed examples - especially in syntactic and semantic domains - are unreliable and inconsistent... Moreover, theoretical linguists are usually unaware of the multiple variables that are known to affect linguistic judgments and can hardly control for them. (Bresnan 2007)

> First, such judgments are inherently unreliable because of their unavoidable meta-cognitive overtones, because grammaticality is better described as a graded quantity, and for a host of other reasons. (Edelman and Christiansen 2003)

In fact, even linguistic methodology seems to acknowledge the cognitive cost of attempting to find an appropriate context for an acceptability judgment: the Truth-

Value Judgment task was created to ease that very cognitive burden such that children are able to give fairly complex linguistic judgments (Crain and Thornton 1998).

At a methodological level, It is obvious that the first issue facing any research that intends to use acceptability judgments as primary data must determine the effect, if any, of context on acceptability judgments. The answer to that question must be determined empirically on a case by case basis. However, once a body of results is gathered, a much more difficult question emerges: What, if anything, does an effect of context tell us about the nature or content of grammatical knowledge?

One possibility, explored in Keller 2000, is that grammatical knowledge is divided into constraints that are affected by context (context-dependent) and constraints that are not affected by context (context-independent). Keller argues that context-in/dependence correlates with a second distinction he draws between *hard* and *soft* constraints, a property that distinguishes between constraints that are present in every language cross-linguistically, and constraints that may or may not be present cross-linguistically. While this is an interesting proposal in its own right, it does not easily transfer from the OT Syntax framework in which Keller is working into other grammatical theories. In other frameworks, the constraints that Keller argues are context-dependent (soft) are usually analyzed as semantic or pragmatic constraints on the use of a given syntactic structure, or preferences for a given interpretation, but not as constraints on the syntactic structure itself. For instance, one of the context-dependent constraints Keller presents is a constraint on gapping from Kuno 1976 that when the remnant material is an NP and VP, they must be interpreted as a standard subject-predicate sentential pattern:

(1) The Tendency for Subject-Predicate Interpretation

a. John persuaded Dr. Thomas to examine Jane and Bill Martha.

b. John persuaded Dr. Thomas to examine Jane and John persuaded Bill to examine Martha.

c. * John persuaded Dr. Thomas to examine Jane and Bill persuaded Dr. Thomas to examine Martha.

In many non-OT frameworks, the fact that this tendency can be overridden by an appropriate context would be interpreted as a pragmatic or information structure effect, with the interesting question then being why the information structure of the gapped sentence presented without any context favors one interpretation over the other, equally structurally possible interpretation. This is exactly the tack taken by Erteschik-Shir 2006.

This second possibility, as proposed by Erteschik-Shir 2006, is that an effect of context indicates an interaction of Information Structure (IS) and the interpretive possibilities of the structure in question. In other words, context-dependent constraints are constraints on Information Structure (IS) not constraints on grammar. For Erteschik-Shir, grammatical constraints yield a binary output, grammatical or ungrammatical, which cannot be altered based on linguistic context. Information Structure constraints, on the other hand, are by definition dependent on the linguistic context as the appropriate information structure changes depending on the discourse function of the sentence. This position is interesting because it transforms context effects into a diagnostic for grammatical knowledge of a sort not normally

investigated by syntactic theory. So in essence, if a structure is found acceptable in any context, it is a grammatically possible structure, but the contexts in which the structure is unacceptable indicate IS-type constraints that require further research.

A third possibility is that context has an effect on acceptability because the complexity of the structure makes the intended meaning difficult to determine. Because acceptability is dependent on the meaning, the grammaticality of the structure is obscured by the indeterminacy of the meaning. Of course, there are many different potential sources for such meaning-obscuring complexity - default processing strategies, default IS strategies, et cetera. - which potentially overlap with the IS proposal of Erteschik-Shir.

Any of these analyses of context effects are possible, and can really only be determined empirically on a case by case basis. The rest of this chapter looks at potential cases of these context effects on properties of wh-movement. Wh-movement, and in particular Island effects, form the empirical object of study throughout the rest of this dissertation, so this chapter serves a dual function: to investigate the possibility that one or more of the scenarios above affects the acceptability of wh-movement structures, and to determine whether context is a factor that must be controlled in future studies in this dissertation. Three studies will be presented:

Experiment 1: The first experiment asks investigates whether the acceptability of Island effects is affected by the complexity of the meaning. In concrete terms, the experiment presents participants with Island violations without any context, and Island violations with a context sentence that is a possible answer to the

question, to determine if fore-knowledge of the meaning of the Island violation affects the acceptability.

Experiment 2: The second experiment follows up on the first experiment in investigating a claim from Deane 1991 that the acceptability of wh-questions is affected by the number of focused elements in the sentence: moved wh-words are interpreted as focused by default, therefore if another element is also focused, which Deane argues is the case for most Island structures as well as some non-Island structures, the two focused elements compete for attention and acceptability is decreased. This experiment investigates the non-Island structures from Deane 1991, and whether conflicting attention is affected by an appropriate discourse context.

Experiment 3: The final experiment takes up Erteschik-Shir's question of whether IS properties are affecting the acceptability of wh-questions when they are presented in isolation. This experiment investigates whether the properties that have been associated with Discourse Linked wh-phrases, such as Superiority amelioration (Pesetsky 1987) and better resumption binding (Frazier and Clifton 2002), arise in non-Discourse-Linked wh-words in an appropriate context.

The picture that emerges from these three experiments is striking: the contexts in these experiments do not interact with the properties of wh-movement. This suggests that the properties of wh-movement under investigation, namely Island effects and Discourse Linking effects, are not due to any of the possibilities discussed above.

In fact, experiment 2 suggests an even stronger result. Deane's analysis of Island effects as instances of conflicting attention cannot be supported by the relative unacceptability of non-Island wh-questions with conflicting attention: experiment 2 found no effect of conflicting attention using sentences taken directly from Deane's paper, regardless of context.

3.2 Island effects and knowledge of the intended meaning

The first experiment asks the simple question: Does fore-knowledge of the intended meaning of an Island violation increase the acceptability? Of course, there are two possibilities. First, knowledge of the intended meaning could increase acceptability of every wh-question equally, leading to an increase in acceptability for grammatical wh-questions as well as Island violations. On the other hand, knowledge of the intended meaning could affect Island violations disproportionately, indicating that a significant portion of the unacceptability of Island violations is due to the difficulty of determining their intended meaning. These two scenarios cannot be distinguished using a simple comparison of the acceptability of Islands to non-Islands, therefore the definition of an Island effect in this experiment will be based on 4 conditions in an interaction of two factors.

The interaction definition of Island effects is easiest to explain with a concrete example. It has been claimed (since at least Huang 1982) that clausal adjuncts are syntactic Islands to wh-movement. This can be demonstrated as an interaction between two factors STRUCTURE and MOVEMENT. STRUCTURE has two levels: clausal

complement (2) and clausal adjunct (3); MOVEMENT also has two levels: movement from the matrix clause (i) and movement from the embedded clause (ii):

(2) Clausal Complement

i. Who$_1$ t$_1$ suspects [$_{CP}$ that you left the keys in the car?]

ii. What$_1$ do you suspect [$_{CP}$ that you left t$_1$ in the car?]

(3) Clausal Adjunct

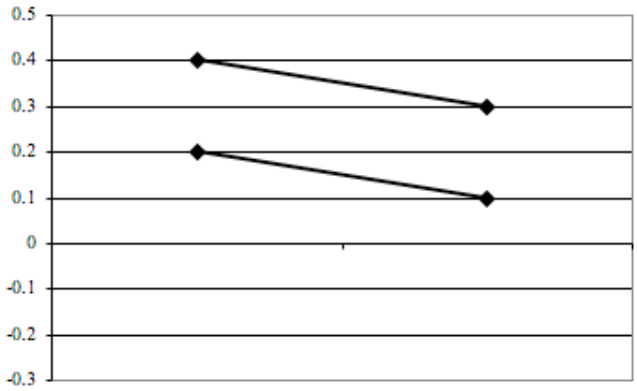
i. Who$_1$ t$_1$ worries [$_{ADJ}$ that you leave the keys in the car?]

ii. * What$_1$ do you worry [$_{ADJ}$ if you leave t$_1$ in the car?]

All things being equal, one might expect sentences containing clausal complements to be judged as more acceptable than sentences containing clausal adjuncts since the semantics of clausal adjuncts involve a more complicated relationship with the matrix predicate. Thus, we might expect both of the examples in (2) to be more acceptable than the examples in (3). We might also expect movement out of the matrix clause to be more acceptable than movement out of the embedded clause (perhaps because shorter movements require less working memory, cf. Phillips et al. 2005), therefore the (i) examples should be more acceptable than the (ii) examples.

If these hypotheses were to hold, we would expect a graph of these four conditions to yield two parallel lines, indicating that there were just two main effects. The slope of the top line shows the effect of distance on wh-movement in sentences containing clausal complements. Similarly, the slope of the bottom line shows the effect of distance on wh-movement in sentences containing clausal adjuncts. The vertical distance between the pairs of points shows the effect of clausal complements versus

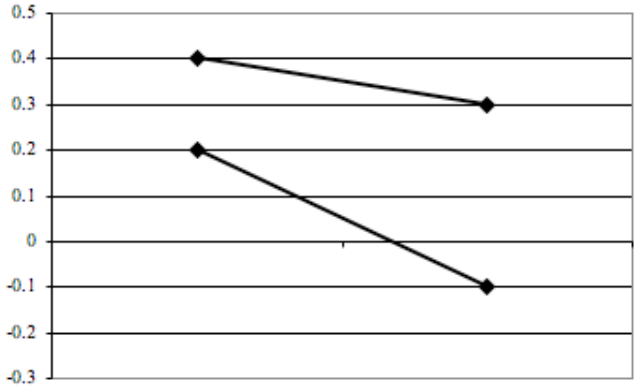
Figure 3.1. Two main effects, no interaction, no Island effect



the effect of clausal adjuncts. This is the standard graph of two main effects (the main effect of STRUCTURE and the main effect of MOVEMENT, and no interaction.

All things are not equal. The claim is that wh-movement out of clausal adjuncts is impossible, because clausal adjuncts are Islands to wh-movement (eg. Huang 1982). This means that there is more affecting the acceptability of the adjunct-embedded condition than just the acceptability decreases we expect from long distance movement and from clausal adjuncts. On the graph, this extra acceptability

Figure 3.2. Interaction, Island effect



decrease would change the slope of the bottom line such that the two lines are no longer parallel. This is the standard graph of a (monotonic) interaction between two factors, and captures the effect of the Adjunct Island while controlling for the contribution of both semantics and wh-movement distance to acceptability.

Nearly every other syntactic Island can be defined through an interaction of STRUCTURE and MOVEMENT:

(4) CNPC Island

i. CP Complement

i. Who$_1$ t$_1$ denied [$_{CP}$ that you could afford the house?]

ii. What$_1$ did you deny [$_{CP}$ that you could afford t$_1$?]

ii. NP Complement

i. Who$_1$ t$_1$ denied [$_{NP}$ the fact that you could afford the house?]

ii. * What$_1$ did you deny [$_{NP}$ the fact that you could afford t$_1$?]

(5) Relative Clause Island

i. CP Complement

i. Who$_1$ t$_1$ knows [$_{CP}$ that the woman read a book?

ii. What$_1$ did you know [$_{CP}$ that the woman read t$_1$?

ii. NP Complement

i. Who$_1$ t$_1$ knows [$_{NP}$ the woman that read a book?]

ii. * What$_1$ do you know [$_{NP}$ the woman that readt$_1$?]

(6) Whether Island

i. CP Complement

i. Who$_1$ t$_1$ thinks [$_{CP}$ that you wrote the letter?]

ii. What$_1$ do you think [$_{CP}$ that you wrote t$_1$?]

ii. Whether Complement

i. Who$_1$ t$_1$ wonders [$_Q$ whether you wrote the letter?]

ii. * What$_1$ do you wonder [$_Q$ whether you wrote t$_1$?]

(7) WH Island

i. CP Complement

i. Who$_1$ t$_1$ thinks [$_{CP}$ that the doctor bought flowers for the nurse?]

ii. Who$_1$ do you think [$_{CP}$ that the doctor bought flowers for t$_1$?]

ii. WH Complement

i. Who$_2$ t$_2$ wonders [$_Q$ what$_1$ the doctor bought t$_1$ for the nurse?]

ii. * Who$_2$ do you wonder [$_Q$ what$_1$ the doctor bought t$_1$ for t$_2$?]

The only minor exception is Subject Islands, for which the two levels of STRUCTURE are simple NPs such as *the manager* versus complex NPs (NPs that contain another NP) such as *the manager of the store*, and the two levels of MOVEMENT are movement out of the embedded object position and movement out of the embedded subject position:

(8) Subject Island

i. Simple NPs

i. What$_1$ do you think the speech interrupted t$_1$?

ii. What$_1$ do you think t$_1$interrupted the TV show?

ii. Complex NPs

i. What$_1$ do you think [the speech by the president] interrupted [the TV show about t$_1$]?

ii. * Who$_1$ do you think [the speech by t$_1$] interrupted [the TV show about whales]?

These 6 sets of conditions will be presented with and without context sentences (sentences which are fully lexicalized answers for the questions) to determine whether fore-knowledge of the meaning interacts with each these Island effects, and if so, to what extent.

Participants

23 Princeton University undergraduates participated in the experiment. All were self-reported native speakers of English with no formal training in linguistics. They all volunteered their time.

Materials and Design

As discussed above, Island effects were defined as the interaction between the two-level factors STRUCTURE and MOVEMENT, yielding 4 conditions for each of 6 islands, or 24 total conditions. 8 lexicalizations of each of condition were constructed and distributed among 8 lists in a Latin Square design. 3 orders of each list were

created (pseudorandomized such that no two related conditions were consecutive), yielding 24 lists total.

Crucially, a third factor, CONTEXT, also with two levels (no context and context), was added to determine whether any of the 6 Island effects are affected by the intended meaning. The context sentence was a fully lexicalized answer appropriate to the target question, with target question to be judged marked in bold:

(9) You think the speech by the president interrupted the TV show about whales.

Who do you think the speech by interrupted the TV show about whales?

The two levels of CONTEXT created two versions of each of the 24 lists, one with context and one without. These two versions were paired to create 24 lists with two sections each (48 total items to be judged). The order of the two sections were counterbalanced, with 12 lists presenting items with context first, and 12 lists presenting items without context first.

The task was magnitude estimation with an If Island reference sentence: *What do you wonder if your mother bought for your father?* The directions were a modified version of the instructions bundled with the WebExp online experimental software suite (Keller et al. 1998). Each section had its own set of instructions, so that the context sentence could be explained. For the context conditions, the reference sentence was also preceded by a context sentence.

3.2.1 Results

Following the standard analysis of magnitude estimation data presented in chapter 2, all of the scores were divided by the reference sentence and log-transformed prior to analysis.

Table 3.1. Islands: descriptive results without context

movement:	matrix				embedded			
structure:	non-island		island		non-island		island	
	mean	SD	mean	SD	mean	SD	mean	SD
Adjunct	.32	.22	.22	.19	.24	.23	-.16	.36
Subject	.16	.25	-.07	.25	.36	.21	-.22	.49
CNPC	.38	.20	.32	.20	.25	.20	-.11	.25
Rel. Clause	.31	.21	.23	.22	.13	.26	-.21	.45
Whether	.38	.19	.29	.18	.24	.22	-.03	.30
Wh	.30	.17	.23	.13	.22	.25	-.23	.37

Table 3.2. Islands: descriptive results with context

movement:	matrix				embedded			
structure:	non-island		island		non-island		island	
	mean	SD	mean	SD	mean	SD	mean	SD
Adjunct	.34	.19	.33	.20	.22	.18	-.16	.16
Subject	.26	.25	-.08	.27	.33	.26	-.25	.20
CNPC	.37	.27	.32	.21	.28	.19	-.10	.23
Rel. Clause	.28	.22	.31	.30	.12	.19	-.22	.33
Whether	.37	.19	.32	.22	.27	.20	.07	.19
Wh	.37	.19	.25	.25	.28	.21	-.17	.24

A three-way repeated measures ANOVA was performed on the factors STRUCTURE x MOVEMENT x CONTEXT for each of the 6 Island types. Island effects are defined as the two-way interaction: STRUCTURE x MOVEMENT, context-dependence as the three-way interaction: STRUCTURE x MOVEMENT x CONTEXT. The results and effect sizes[1] for each Island are given in the following tables:

[1]Partial Eta-squared is a measure of the proportion of variance accounted for by the effect. For

Table 3.3. Results for three-way repeated measures ANOVA

	Adjunct	Subject	CNPC	Rel. Cl.	Whether	WH
STRUCTURE	$***$	$***$	$***$	$***$	$***$	$***$
MOVEMENT	$***$	.730	$***$	$***$	$***$	$***$
CONTEXT	.400	.757	.805	.850	.191	$*$
STRUC X MOVE	$***$	$***$	$***$	$***$	$***$	$***$
CON X STRUC	.164	.482	.989	.324	.070	.623
CON X MOVE	.183	.276	.508	.564	.320	.725
C X S X M	.370	.426	.694	.378	.589	.572

$*** = p < .001, ** = p < .01, * = p < .05$

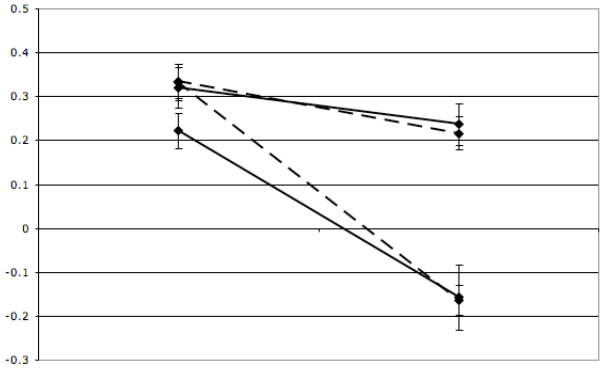
Table 3.4. Partial Eta-squared effect sizes (Cohen 1973)

	Adjunct	Subject	CNPC	Rel. Cl.	Whether	WH
STRUCTURE	.791	.879	.833	.432	.655	.840
MOVEMENT	.826	—	.803	.748	.683	.792
CONTEXT	—	—	—	—	—	.173
STRUC X MOVE	.715	.521	.816	.534	.385	.592
CON X STRUC	—	—	—	—	—	—
CON X MOVE	—	—	—	—	—	—
C X S X M	—	—	—	—	—	—

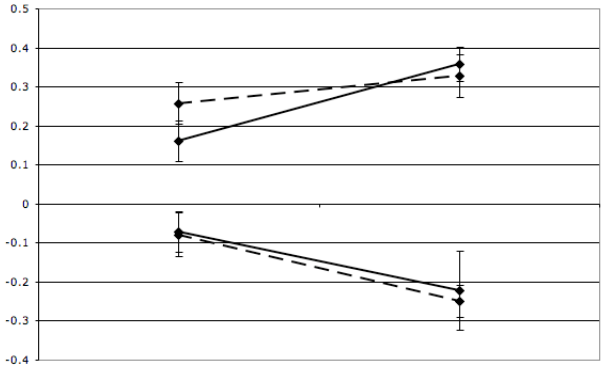
As hypothesized, there are highly significant and very large effects for STRUCTURE, MOVEMENT, and the interaction STRUCTURE X MOVEMENT for all of the islands except for Subject Islands, which do not show an effect of MOVEMENT. This interaction can be seen graphically by the non-parallel lines that emerge when the four conditions of each Island effect are plotted (solid black lines indicate no-context, dashed lines indicate context): The exception to the distance effect for Subject Islands is unsurprising given that MOVEMENT in the Subject Island condition is the difference between movement out of subject position and object position in the same clause,

instance, .840 would indicate that 84% of the variance is accounted for by that effect. Following convention, .01 is considered a small effect, .09 a medium effect, and .25 a large effect

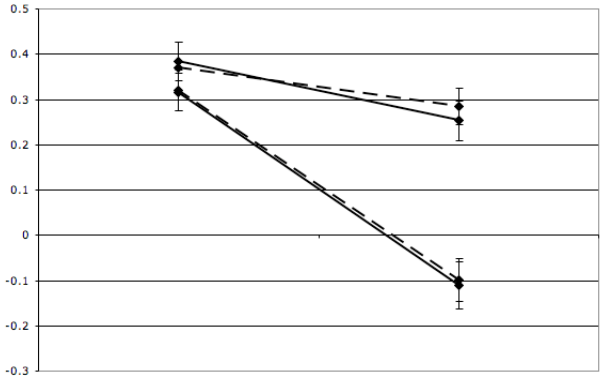
Figure 3.3. Island effects and context



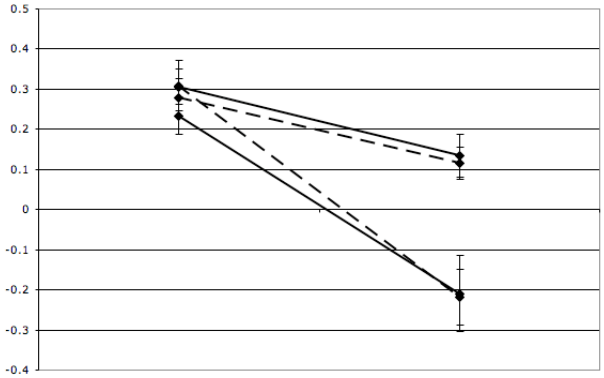
(a) Adjunct Island



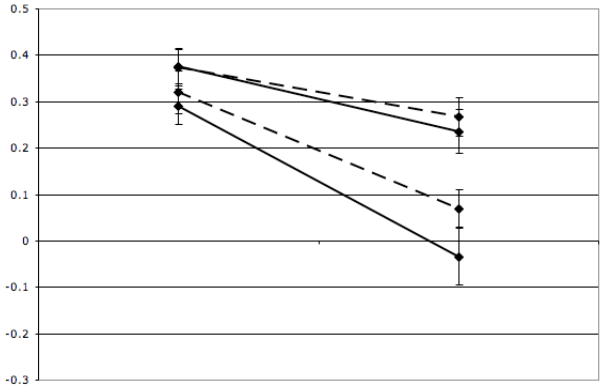
(b) Subject Island



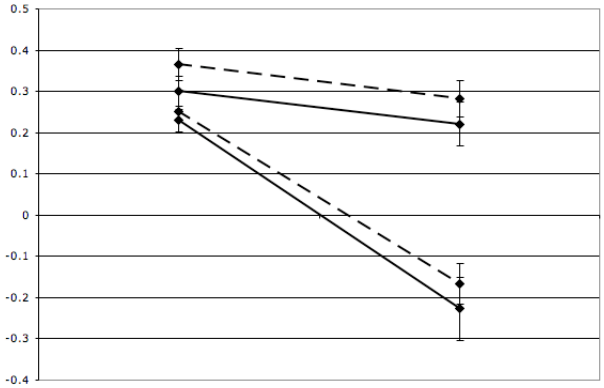
(c) CNPC Island



(d) Relative Clause Island



(e) Whether Island



(f) WH Island

whereas MOVEMENT in the other Island effects is the difference between movement out of the matrix and embedded clauses:

(10) Levels of DISTANCE for Subject Islands

i. What$_1$ do you think the speech interrupted t$_1$?

ii. What$_1$ do you think t$_1$interrupted the TV show?

(11) Levels of DISTANCE for other Islands

i. Who$_1$ t$_1$ thinks [$_{CP}$ that you wrote the letter?]

ii. What$_1$ do you think [$_{CP}$ that you wrote t$_1$?]

The only other effect is the medium-sized significant main effect of CONTEXT on WH Islands. This too is unsurprising given that WH Islands are the only Islands that involve two instances of wh-movement. With two instances of wh-movement and no context sentence, the participants must determine which wh-word goes in which position. The context sentence supplies this information, making the processing of the sentence slightly easier, which may lead to an increase in acceptability.

Discussion

The interaction of STRUCTURE x MOVEMENT confirms that each of these 6 Island effects do indeed exist. However, there is no interaction of CONTEXT with STRUCTURE and MOVEMENT, as can be seen by the fact that the dashed lines in the graphs track the solid lines almost perfectly. This suggests that these 6 Island effects are not affected by the intended meaning. In fact, there is no effect of context in this experiment except for the one main effect on WH Islands, which as previously mentioned, may simply be due to the unique problem of identifying the gap position

of each wh-word in WH Islands.

3.3 Conflicting attention and non-Island unacceptability

The goal of Deane 1991 is to demonstrate that the classic Subjacency account of Island effects (in particular, the formulation in Chomsky 1973) is insufficient to account for the full range of acceptability facts, and that an attention-based account is empirically superior. The argument has two parts: i) there are acceptable sentences that Subjacency predicts should be unacceptable, and ii) there are unacceptable sentences that Subjacency predicts should be acceptable:

(12) Acceptable sentences that should be unacceptable

a. This is one newspaper that the editor exercises strict control over the publication of.

b. Which laws do you advocate an end to the enforcement of?

(13) Unacceptable sentences that should be acceptable

a. * Which car did you like the girls in?

b. * Which crate did you buy the furniture in?

Deane argues that the crucial factor in each of these cases is semantic, and that a theory of attention can capture these facts more adequately than a structural constraint like Subjacency. In particular, Deane argues that unacceptable wh-movement arises when the displaced element must compete for focal attention with another element in the sentence. Whether another element requires attention or not is a consequence of the meaning: for instance, it is not surprising to talk about editors exercising control

over the publication of newspapers, so only newspaper needs attention; however, cars are not often distinguished based upon their female occupants, so *girls* commands attention at the same time as *which car*, leading to unacceptability.

Given that this effect is predicated upon the attention of the participant, and given that attention is determined by how surprising the meaning of the sentence is, this analysis predicts that the effect may be neutralized in an appropriate context. If participants are given a context that biases them to expect the non-moved elements in the sentence, these elements should no longer compete for attention with the moved wh-word, and one might predict that the effect would disappear. The experiment reported in this section tests this hypothesis.

Participants

24 University of Maryland undergraduates, all native speakers of English with no formal training in linguistics, participated in this experiment. All participants volunteered their time.

Materials and Design

This experiment tested 8 pairs of sentences taken directly from Deane 1991. Each pair consisted of 1 sentence with elements that do not command attention, and one sentence with unexpected elements that do command attention, so the pairs formed the factor ATTENTION with two levels: no special attention (acceptable) and special attention (unacceptable). Minor modifications were made to a few of the sentences to make the conditions closer to minimal pairs (for instance, *what type* was changed to *which type* in the first pair): The 8 items in each level were distributed using a Latin Square design yielding 8 non-matched sentence pairs. These pairs were

Table 3.5. Pairs from Deane 1991

acceptable	Which apartments do we have security keys to?
unacceptable	Which type of security key do you have an apartment with?
acceptable	Which reserve divisions do you know the secret locations of?
unacceptable	Which locations do you have reserve divisions in?
acceptable	Which models did you notice shapely legs on?
unacceptable	Which type of legs do you want the models to have?
acceptable	Which wines did you enojy the flavor of?
unacceptable	Which flavor do you want your wine to have?
acceptable	Which books did you enjoy the contents of?
unacceptable	Which type of content do you like your books to have?
acceptable	Which people have you forgotten the names of?
unacceptable	Which names do you know people with?
acceptable	Which store did you buy the furniture in?
unacceptable	Which mood did you buy the furniture in?
acceptable	Which car did you like the gears in?
unacceptable	Which car did you like the girls in?

combined with 3 filler items in a pseudorandomized order such that the two target sentences were not consecutive, yielding 8 lists of 5 items. The lists were then paired with non-identical lists to yield 8 surveys, each consisting of two sections, with each section being 5 items long.

Short (4-5 line) stories were created for each sentence to create a context in which the elements in the unacceptable conditions would be expected. The sentence to be judged was always the final sentence of the story, and was bold. For instance:

> After test driving many cars, the teenager finally came to a decision about which car to purchase. He said to his girlfriend: You know, all of the cars

were really nice to drive, but some of them just felt cooler. Knowing him well, his girlfriend responded: Cooler? I know whats going on. Youve seen commercials for these cars, and some of them had just what you were looking for, like pretty girls driving them. So tell me: **Which car did you like the girls in?**

Two versions of each of the 8 surveys were created: 1 version in which context stories were added to the first section, and 1 version in which the context stories were added to the second section, thus the factor CONTEXT was counterbalanced for order of presentation across 16 total surveys. An 8 item practice section was added to the beginning of each survey consisting of items that cover the full 7 point scale.

The length of the context stories dictated two design decisions. First, as already mentioned, the survey itself was very short: including practice items there were only 18 items in the survey. Second, the task chosen was a 7 point ordinal scale task rather than magnitude estimation: because magnitude estimation involves the comparison of two sentences, in the context conditions it would require two stories, and thus make the task almost unmanageable. The instructions for this task were a modified form of the instructions for previous experiments that included explicit instruction in the 7 point scale task (with two example items at either end of the scale, 1 and 7).

Results

The design of this experiment is a straightforward 2x2 repeated measures factorial design (ATTENTION X CONTEXT), which would normally be analyzed by a

Table 3.6. Mean values and standard deviations for each condition

	Mean	Standard Deviation
context, acceptable	6.12	0.90
context, unacceptable	5.71	1.60
no context, acceptable	5.42	1.53
no context, unacceptable	5.00	2.04

standard two-way repeated measures ANOVA. However, ANOVA assumes normally distributed interval level data, and the 7 point scale task in this experiment only yields ordinal level data which may or may not be normally distributed. Unfortunately, there is no generally accepted non-parametric version of factorial ANOVA, and in fact, standard ANOVA is often reported for ordinal data in the psychological literature. To be safe, two analyses were performed. First, a two-way repeated measures ANOVA was performed on the ordinal data despite violating the assumptions of the test, as appears to be customary within the psychological literature. No significant effects were found with this analysis, but because this could be due to the inappropriate use of ANOVA, a second analysis was performed. The second analysis was a standard two-way repeated measures ANOVA, but it was performed on the rank-transformed version of the ordinal data,[2] following the suggestion of Conover and Iman 1981. As Seaman et al. 1994 point out, the Conover and Iman method is more susceptible to Type I errors - that is, more likely to produce a significant effect. Given that no significant effects were found with this method either, we can be fairly certain that there are no significant effects in the ordinal data:

[2]Rank transformation involves ordering all of the data from lowest to highest, and assigning an order rank to each response

Table 3.7. Two-way repeated measures ANOVAs using ordinal and rank data

	Standard ordinal data		**Rank-transformed data**	
	F	p	F	p
ATTENTION	2.92	.101	.524	.477
CONTEXT	3.85	.062	3.41	.078
ATTENTION x CONTEXT	.000	1.00	.000	.997

There were no significant effects in either analysis, not even of ATTENTION, although there was a nearly significant main effect of CONTEXT which would reach significance under a one-tailed test (if we hypothesized that context always leads to higher acceptability).

Discussion

In the end, the Deane 1991 contrast was not a good candidate for testing the context hypothesis because the crucial contrast either does not exist, or is not detectable by a 7 point scale task. In fact, while it is possible that a more sensitive task such as magnitude estimation may detect a significant effect for ATTENTION, there is reason to believe that the contrast is not one between acceptable and unacceptable sentences, but rather between two acceptable sentences: if one inspects the mean ordinal rankings, we see that all of the conditions are above 5, which is well above the middle of the scale (3.5). This combined with the lack of any significant differences suggests that participants found all of the items acceptable. While this is disappointing from the point of view of investigating the effect of context on Deane's attention-based theory, it is interesting from the point of view of investigating Island effects, as it suggests that the data underlying Deane's analysis is not robust, and therefore not an argument against structural theories such as Subjacency.

3.4 Discourse Linking and context

The final experiment in this chapter investigates whether the properties of Discourse Linking (D-linking) can be bestowed upon non-Discourse-Linked wh-words in an appropriate context, or in other words, whether the properties of D-linking may be true discourse or IS constraints. D-linking was chosen for two reasons: First, as the name suggests, D-linking has been related to the semantic interpretation of the wh-phrase with respect to a given discourse, thus a priori seems like a good candidate for a discourse or IS property. Second, D-linking has been shown to affect acceptability judgments with respect to two major wh-phenomena: Superiority and Resumption. Although neither property will figure substantively in any of the analyses in the rest of this dissertation, they are both major components of any comprehensive theory of wh-movement.

3.4.1 The Superiority effect and D-linking

The Superiority effect is the decrease in acceptability that has been observed for multiple wh-questions in which a structurally 'lower' (usually defined in terms of c-command) wh-word, for instance the object of a verb, is moved 'across' a structurally 'higher' wh-word, for instance the subject of the sentence:

(14) The standard Superiority effect

a. I wonder who read what.

b. * I wonder what who read.

c. I wonder what John read.

In 14a the subject wh-word *who* moves to the specifier of the embedded CP and the sentence is completely acceptable. In 14b the object wh-word *what* moves to the specifier of the embedded CP and the sentence is unacceptable. In 14c *what* again moves to the specifier of the embedded CP, but does not cross over a 'higher' wh-word, and the sentence is again acceptable.

Pesetsky 1987 observed that the Superiority effect disappears when the wh-words are D-linked wh-phrases such as *which student*:

(15) The Superiority effect disappears with D-linking

a. I wonder which student read which book.

b. I wonder which book which student read.

c. I wonder which book the student read.

While a precise semantic definition of D-linking has remained elusive for the past 20 years, the crucial difference appears to be the difference between the possible sets of answers to D-linked and non-D-linked wh-questions. For instance, the set of possible answers to the embedded question in 14a is (almost) any human being and (almost) any piece of reading material. However, the set of possible answers to the embedded question in 15a is not only restricted to students and books, but to students and books that have been previously mentioned in the discourse (or possibly made salient in some non-linguistic way). And as these examples illustrate, D-linking does not depend upon an answer actually being necessary in the conversation: embedded question in English are not normally (or easily) answered in standard conversations, yet the D-linking effect on Superiority persists.

Given that Superiority appears to be context-dependent in that it is affected by discourse restrictions on the set of possible answers, the first question investigated by this experiment is whether a context that restricts the set of possible answers can cause the Superiority effect to disappear with non-D-linked wh-words. Or in other words, can context D-link non-Dlinked wh-words?

3.4.2 The Resumption effect and D-linking

The Resumption effect is an increase in acceptability that has been reported for Island violations in which the illicit gap position is filled by a pronoun agreeing in person and number with the moved wh-word (Ross 1967):

(16) The standard Resumption effect

a. * What$_1$ did you meet a man that read t$_1$?

b. ? What$_1$ did you meet a man that read it$_1$?

Frazier and Clifton 2002 present a series of experiments demonstrating that D-linked wh-phrases are 'more prominent' antecedents for pronouns than non-D-linked wh-words. One of the experiments is a standard 7 point scale acceptability task comparing If Islands with resumptive pronouns and non-D-linked wh-words to WH Islands with resumptive pronouns and D-linked wh-words:

(17) Frazier and Clifton conditions

a. Who did the teacher wonder if they had gone to the library?

b. Which students did the teacher wonder if they had gone to the library?

There is significant effect of D-linking, with mean ratings of 5.58 and 4.87 respectively (1 = good, 7 = terrible) indicating that D-linked wh-antecedent for resumptive pronouns increase the overall acceptability of the sentence. The obvious follow-up question then is whether context can affect the same increase in acceptability by creating a D-linking effect for non-D-linked wh-words. However, there are two confounds in the Frazier and Clifton materials that need to be addressed before manipulating context.

The first confound is in the materials themselves: all of the items tested by Frazier and Clifton involved wh-movement of an embedded subject, or in other words, involved a Comp-trace filter violation. Because the Comp-trace violation existed in all conditions, it would most likely lower the acceptability of all conditions equally. And because the effect is defined as a comparison between two conditions, this should not have affected the results, but could explain why the mean ratings were so low. In order to keep the results of the experiment reported in this section comparable to those of Frazier and Clifton, the Comp-trace violation will be retained in all of the materials constructed for this experiment.

The second confound is the design of the experiment. It has long been observed that D-linking increases the acceptability of Island constraints, especially Whether/If Islands, independently of whether there are resumptive pronouns (e.g., Pesetsky 1987:

(18) The effect of D-linking on Islands

a. * Who did Mary wonder if John had met?

b. ? Which student did Mary wonder if John had met?

This confound means that the effect found by Frazier and Clifton could at least partially be due to the general effect of D-linking on Islands. Given the other (mostly reading time) studies presented by Frazier and Clifton, this does not necessarily call all of their results into question, but it is a confounding factor in their design that should be rectified. As such, the experiment presented here uses a full 2x2 design of the factors resumptive pronouns and gaps, and D-linked and non-D-linked wh-words:

(19) Non-D-Linked

a. Who$_1$ did the teacher wonder if t$_1$had gone to the library?

b. Who$_1$ did the teacher wonder if they$_1$ had gone to the library?

(20) D-Linked

a. Which student$_1$ did the teacher wonder if t$_1$ had gone to the library?

b. Which student$_1$ did the teacher wonder if they$_1$ had gone to the library?

3.4.3 The D-linking experiment

91 University of Maryland undergraduates participated in this experiment for extra credit. All were native speakers of English and were enrolled in an introductory linguistics course. The course did not introduce them to the Superiority effect or the Resumption effect, and the survey was administered in the first half of the semester prior to the discussion of syntactic structure and acceptability judgments.

Materials and Design

There are two sub-designs in this experiment: Superiority and Resumption. The D-linking effect on the Superiority effect is defined as the interaction of the factors

SUPERIORITY and D-LINKING, each with two levels: no Superiority violation (i) and Superiority violation (ii), and non-D-linked 21 and D-linked 22:

(21) Non-D-linked

i. I wish I knew who read what.

ii. I wish I knew what who read.

(22) D-linked

i. I wish I knew which student read which book.

ii. I wish I knew which book which student read.

Similarly, the D-linking effect on the Resumption effect is also defined as the interaction of two factors with two level each: RESUMPTION, with the levels gap (i) and pronoun (ii), and D-LINKING:

(23) Non-D-linked

i. Who$_1$ did the teacher wonder if t$_1$had gone to the library?

ii. Who$_1$ did the teacher wonder if they$_1$ had gone to the library?

(24) D-linked

i. Which student$_1$ did the teacher wonder if t$_1$ had gone to the library?

ii. Which student$_1$ did the teacher wonder if they$_1$ had gone to the library?

The final factor was CONTEXT, yielding two 2 x 2 x 2 sub-designs: SUPERIORITY x D-LINKING x CONTEXT and RESUMPTION x D-LINKING x CONTEXT.

12 lexicalizations of each condition were created, and distributed among 12 lists using a Latin Square design. Context stories were created for each sentence in

which the set of possible answers for the wh-word was restricted. The sentence to be judged was the final sentence of the story, and was bold:

> Last semester, Professor Smith assigned 36 books for his literature students to read, but he knows that no one read all of them. In fact, hes pretty sure that each book was read by only one student, so he wants to only order the books that the literature majors read. Before placing the book order for the next semester, he thinks to himself: **I wish I knew which book which student read. — I wish I knew what who read**.

Two filler items were added to the survey for a total of 10 items, pseudorandomized such that no two related conditions were consecutive, and preceded by 8 practice items spanning the entire range of acceptability. Given the length of the context stories, the task was again the 7 point scale task.

Participants were presented one of the surveys, followed by an unrelated 32 item survey, then followed by a second survey from this experiment. The two surveys were paired such that they did not contain the same lexicalizations. Half of the respondents were given context stories for the first survey but not the second, and half were given context stories for the second by not the first, such that CONTEXT was counterbalanced for order of presentation.

Results

The results of both sub-designs are presented in the following two tables, which report the means and standard deviations for each condition:

Table 3.8. Superiority: descriptive results

context:	no context				context			
D-linking:	non-D-link		D-link		non-D-link		D-link	
	mean	SD	mean	SD	mean	SD	mean	SD
No violation	6.07	1.48	6.23	1.12	6.24	1.41	6.53	1.04
Superiority violation	2.52	1.61	4.78	1.85	3.59	1.69	5.60	1.56

Table 3.9. Resumption: descriptive results

context:	no context				context			
D-linking:	non-D-link		D-link		non-D-link		D-link	
	mean	SD	mean	SD	mean	SD	mean	SD
Gap	1.97	1.14	2.49	1.39	3.02	1.58	3.63	1.52
Resumption	2.58	1.68	3.51	1.65	3.31	1.47	4.53	1.64

Once again, given that the data obtained from the 7 point scale task is ordinal, two analyses were performed: one standard three-way repeated measures ANOVA on the untransformed data, and a second three-way repeated measures ANOVA on the rank-transformed data (Conover and Iman 1981, Seaman et al. 1994). Both analyses were performed on each sub-design:

Table 3.10. Results for three-way repeated measures ANOVA, untransformed data

	Superiority			Resumption		
	F	p	partial-eta^2	F	p	partial-eta^2
SUP/RES	359.8	$***$	.800	46.5	$***$	.341
D-LINKING	125.4	$***$	.582	68.7	$***$	.433
CONTEXT	29.4	$***$	.246	64.1	$***$	.416
SUP/RES X D-LINK	101.2	$***$	.529	9.1	$**$	.092
CON X SUP/RES	16.8	$***$	.158	1.3	.258	.014
CON X D-LINK	0.1	.717	.001	0.8	.365	.009
C X S/R X D	1.4	.238	.015	0.4	.523	.005

As the tables indicate, the two analyses returned the same results. There was a large and highly significant main effect of each factor. There was a large and highly significant interaction of D-linking with the Superiority effect, confirming the

Table 3.11. Results for three-way repeated measures ANOVA, rank-transformed data

	Superiority			Resumption		
	F	p	partial-eta^2	F	p	partial-eta^2
SUP/RES	431.9	$***$	.828	43.5	$***$	.326
D-LINKING	100.4	$***$	.527	77.3	$***$	.462
CONTEXT	24.3	$***$	.213	69.0	$***$	.434
SUP/RES X D-LINK	79.8	$***$	.470	6.4	$*$	.066
CON X SUP/RES	4.0	$*$	.043	2.1	.152	.023
CON X D-LINK	0.7	.403	.008	0.1	.728	.001
C X S/R X D	0.1	.710	.002	0.0	.936	.001

observation of Pesetsky 1987, and a medium sized interaction of D-linking and the Resumption effect, confirming the observation of Frazier and Clifton 2002. There was also a surprising interaction of context and the Superiority effect, but not equivalent interaction of context and the Resumption effect. There was no interaction of context with D-linking or of context with D-linking and Superiority and Resumption.

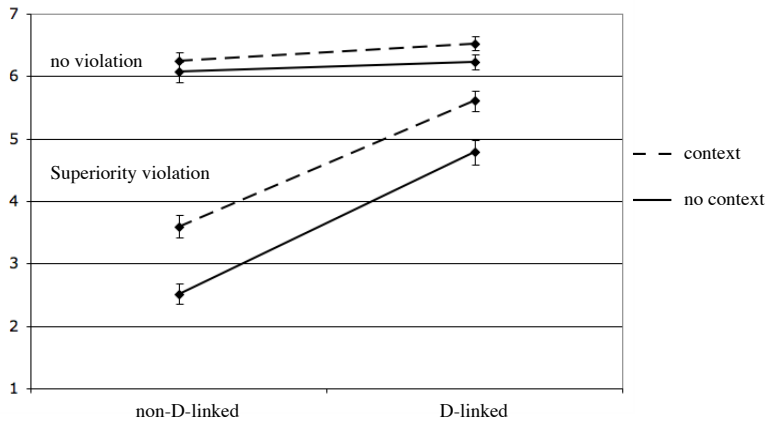
Discussion

Turning to each of the main effects first, we first see that the main effects of SUPERIORITY and RESUMPTION confirm that the Superiority and Resumption effects exist, which is in itself unsurprising. However, we also see that there is a main effect of D-LINKING, suggesting that D-linked wh-phrases increase acceptability even in non-Superiority and non-Island structures. In itself this effect is also not very surprising - D-linked wh-phrases carry more semantic information than non-D-linked wh-words, and acceptability judgments are predicated upon the participants determining the meaning of the sentence, which may be easier with this extra information. However, this effect does raise questions for the observation mentioned briefly before that Island effects are weaker with D-linked wh-phrases. If D-linking increases acceptability for all

structures, then an experiment along the lines of the first experiment in this chapter will be necessary to determine whether D-linking has more of an effect on Islands than non-Islands. The final main effect, CONTEXT, is also unsurprising as we have been amassing evidence throughout this chapter than context increases the acceptability of all structures.

Turning next to the interactions, we see that this experiment confirms the observations of Pesetsky and Frazier and Clifton that D-linking interacts with the Superiority and Resumption effects, as we see an interaction of D-LINKING x SUPERIORITY and D-LINKING x RESUMPTION. We also found a surprising interaction of CONTEXT and EFFECT for Superiority, which suggests that the main effect of context is different for Superiority violations than it is for non-Superiority violations. Graphing the means of each of the conditions highlights this effect, as the vertical distance is greater between the context (dashed lines) and no-context (solid lines) conditions for Superiority violations than it is for no-violation:

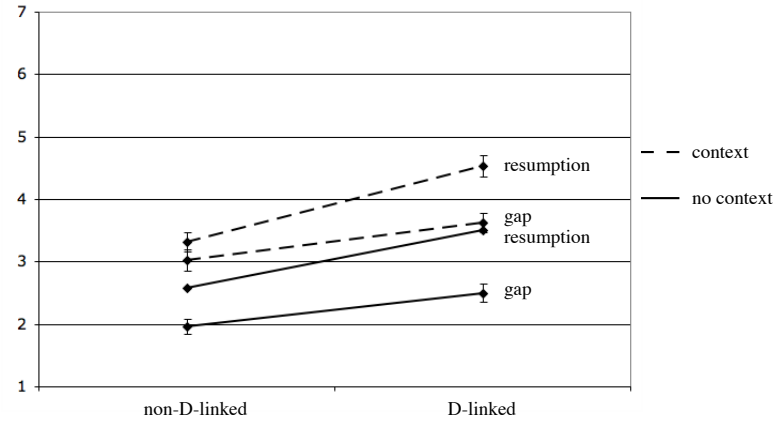
Figure 3.4. Superiority and D-Linking



The fact that context has less of an effect on no-violation may be an artifact of the task: the no-violation conditions in this experiment are already near the ceiling of the scale (with means above 6), and context has a main effect of increasing acceptability. Given that there is very little room left on the scale for increasing acceptability for the no-violation condition, it may be the case that this interaction is really a ceiling effect. A follow-up study using a ceiling-less response scale, such as magnitude estimation, could determine whether this interaction is indeed real or just a ceiling effect of the 7 point scale.

The final two interactions are the two of interest for this experiment, as they represent the ability of context to cause D-linking of non-D-linked wh-words. As the tables report, there is no interaction of CONTEXT and D-LINKING, indicating that the effect of D-linking does not change based on whether there is context or not, or in other words, non-D-linked wh-words do not become D-linked in context. There is also no interaction of CONTEXT, D-LINKING, and SUP/RES, indicating that neither Pesetsky's observation nor Frazier and Clifton's observation change in context, or in other words, non-D-linked wh-words do not become D-linked in context. These findings are reinforced by the parallel patterns of no context (solid lines) and context (dashed lines) conditions in the Superiority graph above, and the Resumption graph below. So it seems once again we've failed to find an interaction of context with properties of wh-movement, although we have (again) found a main effect of context on acceptability judgments in general.

Figure 3.5. Resumption and D-Linking



3.5 The complexity of context and wh-movement

The underlying sources of context effects on acceptability judgments are complex. Compounding this complexity is the fact that the presence and source of context effects must be identified for each piece of grammatical knowledge case by case. This chapter laid out several possibilities for these underlying sources, and investigated several properties of wh-movement to determine whether context affects these properties, and if so, which source causes the effect. The results suggest that various types of context do not affect Island effects or D-linking properties of wh-movement. While this is unfortunate from the point of view of studying context effects, these results are encouraging for the study of structural effects of wh-movement: they suggest that context need not be considered in subsequent acceptability studies. While the effect of context on other aspects of grammatical theory awaits future research, these results suggest, at a minimum, that when it comes to Island effects and D-linking, acceptability judgments taken out of context are robust and stable.

Chapter 4

Satiation and types of grammatical knowledge

This chapter investigates a phenomenon, known in the literature as *syntactic satiation*, in which acceptability judgments of certain violations appear to get better, that is more acceptable, after several repetitions. The fact that some violations satiate while others do not has been interpreted in the literature as an indication of different underlying sources of the unacceptability. In other words, the nature of the violation (or the nature of the grammatical knowledge) affects its relationship with acceptability such that acceptability may change over time. This chapter takes a closer look at the satiation effect using the tool of experimental syntax. Section 1 reviews the existing satiation literature and the motivation for interpreting satiation as an indicator of differences in grammatical knowledge. Section 2 presents a problem for such analyses: the fact that the reported satiation results cannot be replicated. The remainder of the chapter attempts to tease apart various experimental factors that may have led to the satiation effect. The picture that emerges is one in which the satiation effect is actually an artifact of experimental design rather than an natural property of certain violations, suggesting that satiation cannot be interpreted as evidence for different types of grammatical knowledge.

4.1 The problem of *Syntactic Satiation*

Nearly every linguist has been there. After judging several sentences with the same structure over days or even months while working on a project, the acceptability begins to increase - an effect that has come to be called *syntactic satiation* (Snyder 2000). While this sounds like a minor occupational hazard for linguists, it belies a serious problem: the complex analyses created by syntacticians are based upon acceptability judgment data; if that data is unstable, especially if unacceptable sentences tend to become acceptable over time, then there is reason to be skeptical of the analyses. Snyder 2000 offers a provocative response to this state of affairs: if it is the case that some violations satiate while other do not, then this may be a crucial piece of evidence for syntactic analyses. One possibility is that there are different classes of violations, those that satiate and those that do not, which needs to be taken into account by syntactic analyses. Another possibility is that the satiating violations may not be due grammatical effects at all, and may actually indicate that the source of the initial unacceptability is a processing effect (that can be overcome with practice). Whatever the interpretation of satiation, Snyder argues that if it systematically occurs with some violations and not others, than it is not a problem for syntactic analyses at all, but rather a new set of data that needs to be integrated into current analyses.

Snyder 2000 reports an experiment that does indeed suggest that only certain violations satiate. Snyder presented 22 MIT undergraduate native speakers of English with a survey to investigate whether the following 7 violations satiate over

5 repetitions: Adjunct Island, Complex NP Constraint (CNPC) Island, Left Branch Constraint (LBC) violation, Subject Island, That-trace effect, Want-for effect, and Whether Island.

Table 4.1. Violations tested in Snyder 2000

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

The results suggest that Whether Islands and CNPC Islands do satiate, that Subject Islands marginally satiate, and that the other violations do not satiate over 5 repetitions. Prima facie, that only a subset of violations exhibit satiation confirm Snyder's contention that satiation could be a new type of classifying data for linguistic analysis, rather than a problem for previous syntactic analyses.

Interest in these findings has to led to at least two follow-up studies: the first, Hiramatsu 2000, investigates the possibility of using satiation to differentiate natural classes of constraints within the grammar itself, and the second, Goodall 2005, investigates the possibility of satiation to differentiate between grammar-based and processing-based effects. These follow-up studies are near replications of Snyder's original experiment in design, task, and content. However, the results are at best only a partial replication:[1]

[1]Hiramatsu used Snyder's original materials, but added 2 blocks, and therefore 2 instances of each violation to the end of the survey (resulting in 7 instances of each violation) to investigate whether additional exposures would lead to satiation of Subject Islands (which were marginal in

Table 4.2. Summary of results for Snyder 2000, Hiramatsu 2000, and Goodall 2005

	Snyder 2000	Hiramatsu 2000	Goodall 2005
Adjunct Island			
CNPC Island	✓		✓
LBC violation			
Subject Island	(✓)	✓	
That-trace effect		✓	
Want-for effect		✓	N/A
Whether Island	✓	✓	N/A

✓= significant effect, (✓) = marginal effect, N/A = not tested

Given that Snyder's solution to the problem of judgment instability is that satiation is systematic and thus a valid object of study, this lack of replicability is distressing. This chapter continues in the tradition of Snyder 2000 and the follow-up studies, asking whether satiation is in fact a property of certain violations but not others. The picture that emerges is that Snyder's original results are not easily replicable, what I will call the replication problem, suggesting that the source underlying satiation in Snyder's results is not the violation, but some other property of the judgment experiment. A detailed study of the original experiment suggests several aspects of the design that could give rise to the judgment instability, in particular the statistical definition of satiation, the task used, and the composition of the experiment. A series of experiments are conducted to tease apart these factors in an attempt to isolate the conditions necessary to license the type of judgment instabil-

the Snyder 2000 study). Goodall followed the general design of Snyder 2000 in that there were 5 blocks of 10 sentences, but there were 6 violations per block instead of 7. Five of these violations are listed in the table. The sixth violation was the violation of interest, lack of Subject-Aux inversion in structures in which it is obligatory (i.e., non-subject wh-questions).

ity reported by Snyder. The results suggest that judgment instability only arises in unbalanced designs (defined as containing many more unacceptable sentences than acceptable sentences), and is much more likely to occur in categorical tasks such as the yes/no task than in non-categorical tasks such as magnitude estimation. These results suggest that satiation is not a property of judgments or violations in general, but rather an artifact of judgment tasks. This conclusion is further corroborated by a piece of evidence at the center of Snyder's original claim: that in the rare cases when satiation is observed, it tends to be *weak* island violations such as Whether Islands that satiate, not other *weak* violations such as the That-trace effect. This receives a natural explanation under a task-centered account, as it has long been known that the judgment process underlying violations that are easily correctable (e.g., the That-trace effect) is qualitatively different from violations that have no obvious correction (e.g., Island effects) (Crain and Fodor 1987). While these findings cast doubt on Snyder's original solution to the satiation problem (that satiation can be studied like any other property of violations), they simultaneously cast doubt on the satiation problem itself, instead suggesting that judgments are a strikingly stable type of data.

4.2 The replication problem

4.2.1 Confirming the replication problem

Before attempting to identify the factor(s) contributing to the replication problem, the first step is to confirm that the replication problem is more than just an accident of the Hiramatsu and Goodall experiments. This subsection reports three

additional attempts at replication. The first is a direct attempt at replication using the very same materials as Snyder 2000.[2] The second attempt also uses the materials from Snyder 2000, but includes an additional task after each yes/no judgment: participants were also asked to rate their confidence in each yes/no judgment on a scale of 1 to 7. The third attempt uses the same design as that of Snyder 2000, but with a few small modifications and new materials. As we shall see, none of these replications resulted in satiation.

Participants

21 University of Pennsylvania graduate students, native speakers of English, with no formal training in linguistics participated in the direct replication. 21 University of Maryland undergraduates, native speakers of English, with no formal training in linguistics participated in the replication with confidence ratings. 25 University of Maryland undergraduates, native speakers of English, with no formal training in linguistics participated in the modified replication.

Materials and Design

The direct replication was identical in all respects to the study in Snyder 2000. The Snyder 2000 design was a standard blocked design with 5 blocks, each containing 10 items. Of the 10 items in each block, 7 of the items were the violations discussed previously, and the remaining 3 items were grammatical fillers. The order of the items in each block were randomized, and 2 global orders of items were created, each the reverse order of the other. This forward-backward balance for the order of

[2]A special thank you to William Snyder for providing the original materials.

presentation insured that the specific tokens of each violation that were seen 1st and 2nd in one order were 4th and 5th in the other order. Thus, the responses in the first two blocks could be compared to the final two blocks without interference from order of presentation. Participants were presented with 1 sentence to be judged per page. Each sentence was preceded by a context sentence that provided a fully lexicalized potential answer to the question to be judged. The task was a standard yes/no task.

The replication with confidence judgments was also identical in materials and design to the Snyder 2000 study, with the addition of a confidence judgment with every acceptability judgment. Participants were instructed to rate their confidence in their yes/no judgment on a scale from 1 to 7 (1 being the least confident) following each item. Therefore each page of the survey included the context sentence, the question to be judged, a line for indicating yes or no, and a scale from 1 to 7 for indicating confidence in the judgment.

The modified replication followed the general design of Snyder 2000 with a few minor modifications. First, there were 8 violations per block instead of 7, resulting in 2 acceptable sentences per block rather than 3 (the reason for this modification will be discussed in section 2.3). Second, all of the unacceptable sentence types were Island violations. Third, the individual sentences were constructed according to the following parameters: i) the length of all of the sentences was 2 clauses, and the length in number of words was identical for every token of each violation; ii) all of the moved wh-words for the violations were either *who* or *what* to avoid the known acceptability effects of other wh-words[3] (except for LBC violations for which this impossible); iii) all

[3]For instance, wh-phrases involving *which* have been observed to be more acceptable than other

of the names chosen were high frequency (appearing in the top 100 names of the 1980s according to the Social Security Administration). Fourth, the order of the blocks was distributed using a Latin Square design resulting in 5. The island violations tested were Adjunct Island, Coordinate Structure Constraint (CSC), Infinitival Sentential Subject Island (ISS), Left Branch Condition (LBC), Relative Clause Island (RC), Sentential Subject Island (SS), Complex Noun Phrase Constraint (CNPC), and the Whether Island:

Table 4.3. Violations used in the modified replication attempt

Adjunct	What does Jeff do the housework because Cindy injured?
CSC	What did Sarah claim she wrote the article and ?
ISS	What will to admit in public be easier someday?
LBC	How much did Mary saw that you earned money?
RC	What did Sarah meet the mechanic who fixed quickly?
SS	What does that you bought anger the other students?
CNPC	What did you doubt the claim that Jesse invented?
Whether	What do you wonder whether Sharon spilled by accident?

Results

The data from these experiments were analyzed following the procedure in Snyder 2000. The steps to this procedure are discussed in detail in section 2.2.2, so they will not be repeated here. However, the basic method is to compare the number of participants whose judgments changed from *no* to *yes* to those whose judgments changed from *yes* to *no* by using the Sign Test. If the Sign Test returns a significant

wh-words when extracted out of Islands (Pesetsky 1987). Also, wh-adjuncts such as *where, when, how* and *why* can modify most predicates, therefore there is always an acceptable interpretation of Island violations involving wh-adjuncts in which the displaced wh-adjunct modifies the matrix predicate.

result, then that violation is interpreted as satiating. The results for each condition are presented in the following tables:

Table 4.4. Results from the direct replication

	No to Yes	Yes to No	$p - value$
Adjunct Island	5	6	1.0
CNPC Island	2	2	1.0
LBC Violation	0	0	1.0
Subject Island	5	2	.45
That-trace effect	5	3	.73
Want-for effect	2	2	1.0
Whether Island	3	4	1.0

Table 4.5. Results from the replication with confidence

	No to Yes	Yes to No	$p - value$
Adjunct Island	4	4	1.0
CNPC Island	2	0	.50
LBC Violation	1	1	1.0
Subject Island	5	1	.22
That-trace effect	7	2	.18
Want-for effect	5	3	.73
Whether Island	6	5	1.0

Table 4.6. Results from the modified replication

	No to Yes	Yes to No	$p - value$
Adjunct Island	4	1	.38
CNPC Island	3	1	.63
Coordinate Structure Constraint	3	4	1.0
Infinitival Sentential Subject Island	1	1	1.0
LBC Violation	2	3	1.0
Relative Clause Island	3	2	1.0
Sentential Subject Island	3	0	.25
Whether Island	4	4	1.0

Discussion

As one can see, there was no satiation in any of these replications, including the direct replication without any modifications whatsoever. The results can be added to the previous three satiation studies, yielding a new summary of results that confirms that the replication problem is real: no violation satiates in more than 2 studies despite being tested in 5 or 6 identical or nearly identical studies:[4]

Table 4.7. Summary of Snyder 2000 and 5 subsequent replication attempts

	Snyder MIT	Hiramatsu UConn	Goodall UCSD	Direct Penn	Confidence UMD	Modified UMD
Adjunct						
CNPC	✓		✓			
LBC						
Subject	(✓)	✓				
That-trace		✓				N/A
Want-for		✓	N/A			N/A
Whether	✓	✓	N/A			

✓= significant effect, (✓) = marginal effect, N/A = not tested

4.2.2 Deconstructing Snyder 2000

The replication problem suggests that violation-type is not the primary factor in predicting satiation. The question then is whether there are other factors, perhaps components of Snyder's original design, that also contribute to the judgment instability that leads to satiation as defined in Snyder 2000. In any experiment there are three main factors that may contribute to the effect:

[4]The universities from which the sample was taken are included in the following table, as it is sometimes asked whether Snyder's original sample may have been unique in being from a private university that focuses (mainly) on science and technology.

1. the design

2. the task

3. the definition of the effect

The goal of experimental design is to control for any task- or participant-related factors (artifacts) that may contribute to the effect being investigated (Schütze 1996, Cowart 1997, Kaan and Stowe ms). For instance, in these experiments, it is unlikely that fatigue is contributing to changes in judgment for any given sentence, as the order of presentation is balanced across all of the participants (if Participant 1 sees sentence A first, then Participant 2 sees sentence A last). However, these experiments do not control for the possibility that participants are biased in their responses, perhaps due to a response strategy: 70% of the items in the survey are by hypothesis unacceptable, which has the potential to bias participants toward judging sentences as acceptable later in the experiment in an attempt to balance their responses.[5]

The task chosen for these experiments was the *yes/no task* in which participants are asked whether a sentence is an acceptable sentence of English. The yes/no task is a categorization task consisting of two categories, and therefore suffers from two drawbacks given the design used in these studies. First, because there are only two responses, the likelihood of a balanced response strategy increases: it is relatively

[5]One anonymous LI reviewer suggests that the unbalanced nature of the composition ensures that participants are not memorizing their responses. While this is not stated as a goal by Snyder himself (2000), it is possible, and would mean that both unbalanced and balanced designs would introduce an artifact (response strategy versus memorization). This issue is taken up in a later section devoted to controlling for participants that may be memorizing their responses.

easy to track two response types, and as the experiment progresses, realize that one is being used disproportionately more often than the other. In fact, verbal debriefing of participants confirms this as nearly every participant asked why there were so many 'bad' sentences. Second, as a very extreme categorization task, the yes/no task is prone to lose potentially relevant data: there is a growing body of research indicating that participants can differentiate surprisingly many levels of acceptability, suggesting that acceptability may be best characterized as a continuous quantity (e.g., Bard et al. 1996, Keller 2000).

The operational definition of satiation used in these experiments was as follows:

1. Count the number of *yes* responses after the first two exposures of each violation for each participant
2. Count the number of *yes* responses after the last two exposures of each violation for each participant
3. If the number of *yes* responses in the last two exposures is higher than in the first two, the participant is defined as satiating for that violation
4. If the number of *yes* responses in the last two exposures is lower than in the first two, the participant is defined as not-satiating for that violation
5. For each violation, count the number of participants that satiated and the number of participants that not-satiated (n.b., the participants whose responses were stable are ignored)
6. If there are statistically more satiaters than non-satiaters then the violation is said to satiate.

Basically, this definition asks: For those people who have unstable judgment, are more of them unstable in a positive direction or in a negative direction? While this is a possible topic of investigation, because of the elimination of stable participants from the analysis, this definition artificially limits the scope of satiation in two important ways. First, satiation is no longer a property of violations in speakers of English,

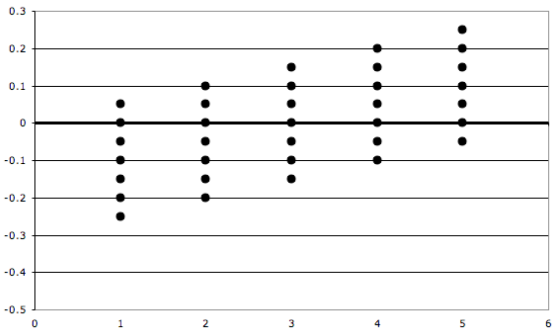
it is a property of violations in speakers of English who have unstable judgments, which is a smaller population - using Snyder's original results as an estimate, for CNPC Islands only 23% of the population is unstable, for Whether Islands 55% of the population is unstable. Second, the question of whether the instability is positive or negative is a biased question: these are violations, which all things being equal, will be more likely to be judged *no* than *yes*. If there is instability at all, then one would expect it to manifest itself as a change from *no* to *yes* because of the initial disproportion.

One can see how these three factors could interact to license the type of instability that is labeled satiation under Snyder's original definition. The task involves two response choices, so participants are likely to employ a strategy to balance them. The disproportionate number of unacceptable sentences means that the strategy will be one that leads to more yes responses later in the experiment. Because stable participants are excluded, the final analysis is conducted over those participants who demonstrated instability, or in other words, participants who are likely to have employed just such a strategy. The fact that these violations are initially judged unacceptable, and the fact that the composition of the survey leads to a strategy of increasing yes responses, make it unsurprising that a satiation effect is found when satiation is defined as comparing the number of participants changing from *no* to *yes* to the number changing from *yes* to *no*.

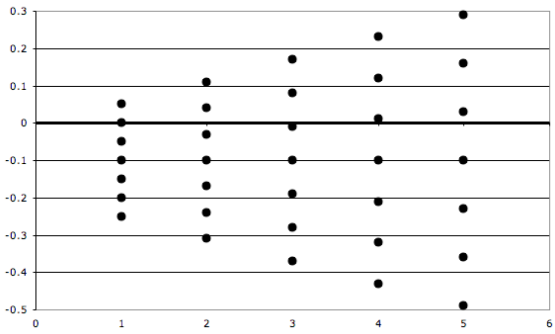
Of course, this limited definition of satiation could be an object worth studying in itself (e.g., *Why does this strategy tend to affect Whether Islands but not That-trace effects?* - a question that is briefly considered at the end of this chapter). However,

given that the effect is defined over the yes/no task, it still is not obvious that the results can be interpreted in a meaningful way. If acceptability is indeed a continuous quantity, then the categories of *yes* and *no* might be masking the true nature of the acceptability judgments. For instance, given a definition in which satiation is an increase in yes responses over time, there are at least three potential models for the actual nature of the acceptability judgments:

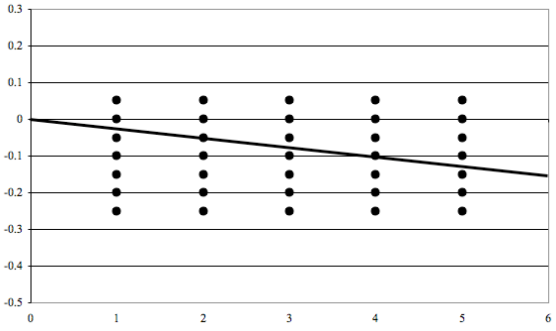
Figure 4.1. Three models of satiation



(a) Mean increase



(b) Variance increase



(c) Threshold decrease

The first model is one in which the acceptability judgments increase over time, and eventually those that started below the yes/no threshold (**solid line**) cross it to become yes responses. This is the model that Snyder and others have assumed is underlying the satiation effect in the previous studies, and correspondingly, is the one

that is informative from the point of view of studying the violations themselves. The second model is one in which the mean acceptability of the violation does not change at all, but the spread or variation of the judgments does change over time, such that some of the judgments that were below the yes/no threshold cross it over time. If this were the model underlying previous satiation effects it would be evidence that judgments are not a good source of data, at least for some violations. The final model demonstrates that defining satiation based on a categorical judgment means that the effect could be the result of a change in the category threshold rather than a change in the underlying percept. If the threshold decreases over time, it would have the same effect: an increase in the number of yes responses. If this were the model underlying previous satiation findings, then it would simply be evidence that categorization tasks do more than lose potentially relevant data, they may also indicate changes in the data when no changes actually occurred.[6]

[6]In the process of teasing apart the factors that contribute to the instability that Snyder interpreted as satiation, these models will begin to be teased apart. Models 1 and 2 are trivial to investigate as they simply require a non-categorical task. Model 3 is probably impossible to measure directly, as the threshold in yes/no tasks is most likely due to a combination of normative factors (grammar, processing, frequency, information structure, context, etc.). However, suggestive evidence for model 3 will come from two sources: i) that models 1 and 2 appear to be incorrect, and ii) that confidence in yes/no judgments decreases over the course of these experiments, which could be caused by a change in the category threshold.

4.2.3 A roadmap

The contribution of these three factors (design, task, and definition) to the instability of judgments can be teased apart through independent manipulation across several experiments. For instance, it is fairly straightforward to cross the factor type of TASK (with 2 levels: yes/no (non-categorical) and MagE (categorical)) with the factor type of DESIGN (also with two factors: balanced and unbalanced), resulting in a standard 2x2 design:

Table 4.8. Crossed design of factors TASK and DESIGN

	Yes/No	MagE
Balanced	?	?
Unbalanced	unstable	?

The problem with the definition of satiation is probably the hardest to manipulate. The Snyder 2000 definition is technically a valid definition, albeit with a very limited scope. To ensure that the experiments in this study are directly comparable to the original studies it will be important to draw a distinction between *satiation*, which is one particular statistical definition (that can be argued over), and *instability*, which is any statistically definable change in judgments for a given structure. Broadening the domain of investigation in this way allows for the possibility that the satiation effect only occurs under very specific circumstances (for instance under certain tasks with certain designs), thus allowing an investigation into the replication problem. I will use the terms *stability/instability* or *stable/unstable* to refer to this change in contrast to *satiation* where appropriate.

4.3 The yes/no task and balanced/unbalanced designs

Crossing the factors TASK and DESIGN leads to 4 cells. One of these cells has been studied extensively: the effect of yes/no tasks in unbalanced designs. As we have seen, these two factors do lead to judgment instability with some violations in some experiment, but not in every experiment. One of the experiments in this cell differed from Snyder's design in two small ways: first, it was composed of 8 violations instead of 7, and second, the violations have all at one point or another been classified as Island violations. The reason for these changes can now be made explicit: If we focus attention on the two violations that Snyder interpreted as satiating (Whether Island and CNPC Island), then this design is really 6 non-satiating violations, 2 satiating violations, and 2 completely acceptable sentences. We can then manipulate the design to be the inverse, while maintaining the two satiating violations as a pivot point: 2 non-satiating violations, 2 satiating violations, and 6 completely acceptable sentences. Thus, the effect of the factor DESIGN on the yes/no task can be isolated:

Table 4.9. Design manipulation for yes/no task

Unbalanced	**Balanced**
Adjunct Island	Acceptable Sentence
Coordinate Structure Constraint	Acceptable Sentence
Infinitival Sentential Subject Island	Acceptable Sentence
LBC Violation	Acceptable Sentence
Relative Clause Island	Acceptable Sentence
Sentential Subject Island	Acceptable Sentence
CNPC Island	*CNPC Island*
Whether Island	*Whether Island*
Acceptable Sentence	Adjunct Island
Acceptable Sentence	Relative Clause Island

Participants

25 University of Maryland undergraduates, all monolingual speakers of English, none with formal exposure to linguistics, participated in the unbalanced experiment. 19 undergraduates participated in the balanced experiment.[7] The experiments were administered in individual testing rooms in the Cognitive Neuroscience of Language Laboratory at the University of Maryland. Participants also completed an unrelated self-paced reading study during their visit to the lab. All of the participants were paid for their participation.

Materials and Design

As already mentioned, these experiments were designed to mimic the design of previous satiation studies while manipulating the balance of unacceptable to acceptable sentences. Therefore the items were divided into 5 blocks of 10 items, with the composition of the 10 items manipulated as outlined above. All of the items were wh-questions in which an argument wh-word (*who* or *what*) is moved, except for LBC violations for which this is impossible. All of the items were controlled for length in clauses and in number of words. The order of presentation of the blocks was distributed using a Latin Square design, and the items within each block were pseudorandomized such that two acceptable sentences did not follow each other in the unbalanced design, and two violations did not follow each other in the balanced

[7]The sample size in Snyder 2000 was 22, which was the target for both of these experiments. Human error in the scheduling of participants resulted in 3 additional participants in the unbalanced experiment, and 3 fewer in the balanced, as these were run concurrently. The unequal sample sizes are compensated for in the statistical analysis.

design. The instructions for the task were identical to those in Snyder 2000.

Results

Applying the Snyder 2000 definition of satiation yields no significant effects in either experiment by Sign Test:

Table 4.10. Results from the unbalanced yes/no task

	No to Yes	Yes to No	$p - value$
Adjunct Island	4	1	.38
Coordinate Structure Constraint	3	4	1.0
Infinitival Sentential Subject Island	1	1	1.0
LBC Violation	2	3	1.0
Relative Clause Island	3	2	1.0
Sentential Subject Island	3	0	.25
CNPC Island	3	1	.63
Whether Island	4	4	1.0

Table 4.11. Results from the balanced yes/no task

	No to Yes	Yes to No	$p - value$
Adjunct Island	0	0	1.0
Relative Clause Island	2	0	.50
CNPC Island	0	0	1.0
Whether Island	0	2	.50

Applying a basic definition of instability, for instance counting the number of participants whose judgments change even once, we find that 15 out of 25 participants' judgments change in the unbalanced experiment, and 3 out of 19 participants' judgments change in the balanced experiment. Fisher's exact test reveals that this difference is significant: $p < .04$ (Fisher 1922).

Discussion

Once again, we face the replication problem: there were no effects by Snyder's definition in either experiment. However, by looking for any statistically definable

instability, in this case, the number of participants that show a change in judgments, we see that there is a significant effect of design: unbalanced designs lead to much more instability. In fact, there was barely any instability under the more balanced design. This is a first step toward our goal of isolating the factors or interactions that lead to judgment instability: at least within yes/no tasks, balanced designs are less susceptible to judgment instability - as expected under the theory that unbalanced designs lead to a response strategy that causes participants to include more yes responses later in the experiment.

4.4 The magnitude estimation task and balanced/unbalanced designs

Recall that there are at least two reasons that yes/no tasks are a less than ideal choice for investigating judgment instability. First, as a categorization task with only two categories, they may be more likely to lead to a response strategy under an unbalanced design because it is fairly straightforward for a participant to track the proportion of two responses. Second, the categorization of responses obscures the true nature of the acceptability judgments, leaving at least three different types of instability that could lead to a satiation effect, and no way to determine which is actually the cause. Overcoming these two problems simply a non-categorization task for measuring acceptability, such as magnitude estimation (Stevens 1957, Bard et al. 1996).

4.4.1 Magnitude estimation and balanced designs

Having seen the effect of balanced and unbalanced designs on the yes/no task, and having seen the benefits of magnitude estimation over categorization tasks, the next logical step is to investigate whether design has a similar effect on magnitude estimation, and if so, where the source of the instability lies. The first set of experiments investigate whether judgments collected using magnitude estimation are stable under a balanced design. In this case, balanced design refers to a set of design properties which are part of the best practices of psycholinguistic experimental design (Kaan and Stowe ms):[8]

1. Ratio of acceptable items to unacceptable items is 1:1

2. Ratio of distracters to experimental items is 2:1 (to minimize strategies)

This subsection reports the results of 5 magnitude estimation experiments using this type of balanced design. Each of the first 4 tested a different Island violation (Subject, Adjunct, Whether, and CNPC Islands). Because the magnitude estimation task requires the comparison of two sentences, context sentences were not included in these 4 designs even though they were part of Snyder's original design (as well as the replication attempts). To ensure that the lack of context sentences had no effect on the results, a fifth experiment was included in which the CNPC Islands were tested

[8]There are undoubtedly many more "best practices" for materials construction (e.g., controlling for the frequency of lexical items). And while the Snyder 2000 materials violated some of these best practices as well, it seems likely that they contributed noise across the conditions, not artifacts into any single condition.

again, but this time with context sentences along the lines of those in Snyder 2000.

Participants

University of Maryland undergraduates with no formal linguistic training participated in these 5 experiments. All were self-reported monolingual speakers of English. Two of the experiments, Subject and Adjunct Islands, were administered over the internet using the WebExp experimental software suite (Keller et al. 1998) in exchange for extra course credit. The other three experiments were conducted in the Cognitive Neuroscience of Language Lab at the University of Maryland, during which participants also participated in an unrelated self-paced reading study and were paid for their time. The sample sizes were 20, 24, 20, 17, and 20 for Subject, Adjunct, Whether, CNPC, and CNPC Islands with context respectively.[9]

Materials and Design

The general design for each of these experiments was identical. Materials were distributed using a blocked design. Each block contained 2 tokens of the island violation, 1 unacceptable distracter, and 3 acceptable distracters. Thus, the composition of each block ensured adherence to the balanced design ratios of 1:1 acceptable to

[9]The sample sizes for Subject and Adjunct Island experiments were based upon responses to an extra credit offer, and therefore were contingent upon the number of students who followed through and completed the survey. Given that the smaller sample size was 20 in these two experiments, the target sample size for the Whether, CNPC and CNPC with context experiments was also 20. However, 3 participants were eliminated from the analysis of the CNPC analysis for reporting a second language spoken in the home.

unacceptable and 2:1 distracters to experimental items:[10]

Table 4.12. Composition of each block in balanced magnitude estimation experiments

sentence type	**judgment type**	**item type**
Island violation	unacceptable	target
Island violation	unacceptable	target
unacceptable sentence	unacceptable	distracter
acceptable sentence	acceptable	distracter
acceptable sentence	acceptable	distracter
acceptable sentence	acceptable	distracter

Both the experimental items and distractors were controlled for length in clauses and length in number of words (with the exact number varying by experiment given the differences among the Island violations). All of the experimental items involved movement of the wh-arguments *who* or *what* to avoid garden-path problems with wh-adjuncts and observed acceptability differences with wh-phrases involving *which* (Pesetsky 1987). Half of the experimental items used the present form *do* and half the past *did*. The distracters included all possible wh-words for

[10]One may wonder why this design was chosen over the balanced design used in the yes/no tasks. The answer is straightforward: because there was no effect (no instability) in that design, it seems unlikely that there would be an effect under magnitude estimation. While running that design under magnitude estimation would certainly prove the stability point that is made with these experiments, the fact that there are only 5 exposures of each violation type in that design means that the results would be of limited value (e.g., perhaps satiation would occur after 7 exposures as Hiramatsu 2000 has claimed for Subject Islands). These designs not only allow us to closely follow general psycholinguistic best practices, but also allow us to increase the number of exposures of each Island type (10 or 14) without overburdening the participants with too many judgments (recall that magnitude estimation requires comparing two judgments for each data point, and a little bit of math, so it is at least twice as taxing as yes/no judgments).

variety (crucially including *who* and *what* to avoid response strategies).

Table 4.13. Exemplars of each of the Islands investigated

Subject	What do you think a movie about would be scary?
Adjunct	Who did you leave the party because Mary danced with?
Whether	What do you wonder whether you forgot?
CNPC	Who do you deny the rumor that you kissed?

The instructions for all of the experiments were a modified version of the instructions published with the WebExp experimental software package. The modifications were i) changing the example items from declaratives to questions because of the nature of the experimental items, and ii) including a short passage indicating that the task was not a memory task to discourage participants from attempting to memorize their responses. All of the experiments were preceded by two practice phases: the first to teach them the magnitude estimation task using line lengths, the second to teach them the task using acceptability.

It should be noted that there were some minor differences among the experiments. However, there are no theoretical reasons to suspect that these differences would affect the result. In fact, the differences were included in an attempt to ensure that the lack of instability (i.e., the stability) found in these experiments was not due to some unknown experimental design decision. Thus, the fact that these differences do not actually result in any effect is further corroboration of the striking stability we shall see in the results section:

(25) Differences among the experiments

i. Subject and Adjunct Islands were 7 blocks long (14 exposures), the others were 5 (10 exposures)

ii. Subject and Adjunct Islands used the reference *What did Kate prevent there from being in the cafeteria?*, the others used the reference *What did you say that Larry bought a shirt and?*. Neither sentence type has ever been claimed to satiate.

iii. The unacceptable distracters for Subject and Adjunct Islands were agreement violations. The unacceptable distracters for the others were Infinitival Sentential Subject Island violations. Neither sentence type has been claimed to satiate.

iv. Subject and Adjunct Islands were administered over the internet, the others in the lab.

Results

Because the unit of measure is the reference value, the first step in analyzing magnitude estimation data is to divide all of the responses by the reference value to obtain a standard interval scale of measure. Because responses are made with the set of positive numbers, and because the set of positive numbers is unbounded to the right and bounded by zero to the left, magnitude estimation data is not normally distributed (it has a rightward skew). To correct for this non-normality, standard practice in the psychophysical literature is to log-transform the responses prior to analysis. The log transformation is chosen for at least two reasons: first, it minimizes

the impact of large numbers, thus bringing the data closer to normal; second, it is straightforward to calculate the geometric mean after a log transformation (it only requires exponentiation), and given that psychophysics is concerned with the ratios of stimulus estimate, the geometric mean is the correct choice of central tendency. In linguistic magnitude estimation, there is no possibility of ratios (no meaningful zero point for acceptability), so the second reason for using the log transformation does not hold. However, it is still the case that the data is non-normal, therefore must be transformed prior to statistical analysis. While there are many transformations available in inferential statistics, the log transformation is well established within the magnitude estimation literature, and ensures that the analysis of linguistic magnitude estimation is only different from psychophysical magnitude estimation because of the stimulus of interest.

Repeated measures linear regressions, following the method proposed in Lorch and Myers 1990, were performed on the means of the log-transformed judgments for each island to determine whether the mean of the responses changed after repeated exposures. The essence of the test is to compare two lines: the first line is simply the horizontal line defined by the grand mean of every response, thus a line that assumes no change based on the number of exposures; the second line is the line of best fit obtained by looking at each exposure to the violation independently. If there is an effect of repeated exposures, then this second line will be significantly different from the grand mean line. Although there is no standard method for reporting linear regression coefficients, a table listing the y-intercept and slope of each line is provided (the two coefficients that define any line, represented in linear regression by

the variable *b* and *Exp(b)* respectively), along with the p-value of of the comparison between this line and the grand mean. As is evident, there is no effect of repetition on the means:

Table 4.14. Linear regressions for means of magnitude estimation in a balanced design

	(b)	Exp(b)	p value
Subject Island	-0.130	0.003	.14
Adjunct Island	-0.320	0.003	.52
Whether Island	0.080	0.008	.14
CNPC Island	0.001	-0.006	.44
CNPC with context	-0.020	0.010	.22

Of course, as demonstrated by model 2 above, it is possible that the satiation effect is found in the variation or spread of the scores over time, not in their means. Therefore a repeated measures linear regression was also performed on the residual scores. Residual scores are the absolute value of the difference between each score and the grand mean, and form the basis for Levene's test for homogeneity of variance (Levene 1960). If the spread of the scores is increasing over time, then there should be an increase in residual scores over time. A table similar to the one for means is reported, and as is evident, there is no effect of repetition on the residuals. A representative scatterplot (Subject Island) of the scores and the non-significant trendline is also included below for a graphical representation of the linear regression method.

Discussion

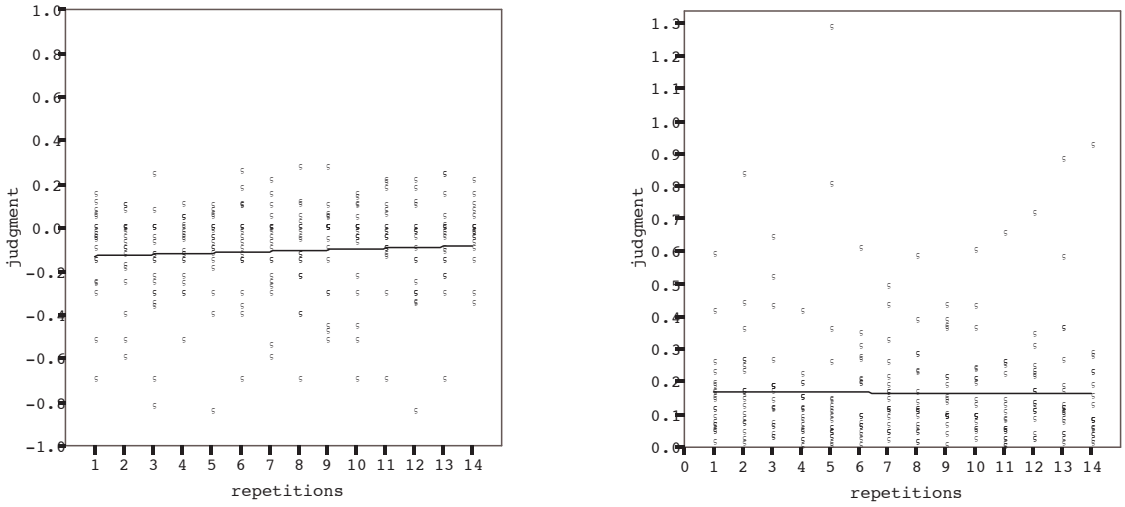
There are three major points made by these analyses. First, there is no increase in mean judgments using the magnitude estimation task in a balanced design. Second, there is no increase in spread of judgments using the magnitude estimation task in a balanced design. And finally, the fact that the first four experiments did not include

Table 4.15. Linear regressions for residuals of magnitude estimation in a balanced design

	(b)	Exp(b)	p value
Subject Island	0.17	-0.001	.83
Adjunct Island	0.26	0.004	.26
Whether Island	0.26	-0.002	.68
CNPC Island	0.27	0.010	.18
CNPC with context	0.35	-0.007	.25

Figure 4.2. Scatterplots and trendlines for Subject Islands

Means **Residuals**



context sentences did not have an effect, as there was no effect of mean or residuals in the CNPC Island with context experiment. These facts are not entirely surprising given the lack of instability of the yes/no task with a balanced design. However, taken together, these results indicate that the instability that lead to the satiation effect in Snyder 2000 does not persist in balanced designs, either for categorical yes/no data or for non-categorical magnitude estimation data.

4.4.2 The memory confound

As an anonymous reviewer correctly points out, the stability we have seen under balanced designs could be due to a confound that is always present in balanced designs: balanced designs increase the likelihood that participants are able to track their responses, and this memorization may lead to the statistical stability. While the instructions explicitly direct participants not to attempt to memorize their responses, it may be a natural phenomenon beyond their control.

In principle, there are two types of evidence that may bear on this problem. First, if participants were indeed memorizing their responses, one might expect them to report this during post-experiment debriefing. One of the standard questions during debriefing is whether they noticed any sentences being repeated, or any sentences that seemed similar to others. In these experiments, no participants reported noticing such similarities. However, given that the participants had no training in linguistics, they may have lacked the vocabulary to describe their intuitions. Indeed, many of the participants were highly cognizant of there being 'bad' sentences since they had never encountered an acceptability task before.

The second piece of evidence of memorization would be identical responses to identical stimuli within each participant. Of course, such consistency could be genuine (the participant could be very good at characterizing acceptability), but to be conservative, a second set of linear regressions were performed after REMOVING any participant that reported the same judgment 5 or more times. Given the nature of memorization, one might choose to only remove participants with repeated scores

later in the course of the experiment, or to only remove participants with consecutive identical judgments. However, to be safe, participants with repeated scores at any point in the experiment, even if none of the repetitions were consecutive, were eliminated. Thus, the participants that remained in the analysis were those with no overt indication of having memorized their judgments.

Linear regressions for the remaining participants follow. There are additional columns to indicate the number of participants who showed significant internal satiation in that their individual judgments showed an increase over time (Satiaters), and the number of participants left in the sample (N) after removing participants that potentially memorized their judgments:

Table 4.16. Linear regressions for means after removing participants

	(b)	Exp(b)	p	Satiaters	N
Subject Island	-0.15	0.005	.11	0	15
Adjunct Island	-0.27	0.006	.41	2	12
Whether Island	0.03	0.010	.19	3	13
CNPC Island	0.03	-0.004	.67	1	9
CNPC with context	0.03	0.005	.65	1	13

As can be seen, judgments were statistically stable even after removing the obviously consistent participants because they could have potentially memorized their judgments. In fact, even looking at each participant individually yields 3 or fewer satiating participants in each sample, well below the critical threshold of 6 required for a potentially significant result using the Sign Test from the satiation definition of Snyder 2000. So it seems clear that memorization is not the cause of the stability seen in the balanced magnitude estimation experiments.

4.4.3 Magnitude estimation and unbalanced designs

The final cell of our crossed design is the stability of data collected using the magnitude estimation technique with an unbalanced design. In this case, the unbalanced design used are the materials from the unbalanced yes/no design in section 2.3. Even though no satiation effect was found for these materials under Snyder's definition using the yes/no task, instability was recorded in the form of participants changing their judgments. Thus, in addition to determining whether there is instability for magnitude estimation and unbalanced designs, this experiment may also bear on the source of the instability found in the yes/no experiment.

Two experiments were conducted using this design. The only difference between the two was the reference sentence used: the first experiment used the Coordinate Structure Constraint (CSC) violation *What did you say Larry bought a shirt and?*, and the second the *If* Island violation *What did you wonder if Larry had bought?*. The logic behind this manipulation is as follows: The CSC-reference was initially chosen because it is an Island-type violation, but has never been claimed to satiate (it would be problematic if the reference sentence changed in acceptability over time). However, the CSC violation is also considered a very strong violation. So while the 8 violations in this design are by hypothesis unacceptable, they are not necessarily worse than a CSC violation, which means that there was no pattern to the whether scores were higher or lower than the reference sentence. While it is not clear whether participants actually track their responses with respect to whether they are higher or lower than the reference, it is at least conceivable that some do.

To ensure that this did not have an affect on the judgments, a second experiment was conducted using an If Island as reference. If Islands are theoretically identical to Whether Islands, therefore are also in the middle range of acceptability. This would ensure that the majority of the violations in the study should be judged worse than the reference. The drawback to this design is that If Islands may be expected to satiate given their relation to Whether Islands (one of the satiating violations in Snyder 2000). A priori, this worry is tempered by the fact that the magnitude estimation experiments in the previous subsection indicate that Whether Islands do not change over time. More interestingly, if it were the case that the If Island reference satiated over time, then the experiment should yield several *negative* effects in which other violations appear to decrease in acceptability, simply because the reference is increasing in acceptability. As we shall see, this was not the case.

Participants

31 University of Maryland undergraduates participated in the CSC-reference experiment, and 22 in the If-reference experiment. All were monolingual, native speakers of English with no formal linguistics training. Participants also participated in an unrelated self-paced reading experiment during their visit to the lab. All participants were paid for their participation.

Materials and Design

The design of these two experiments is identical to the yes/no version: 5 blocks of 10 sentences, with each block containing 8 violations and 2 acceptable distracters. The violation types in each block are repeated below for convenience:

Table 4.17. Violations in unbalanced MagE task

Adjunct Island
Coordinate Structure Constraint
Infinitival Sentential Subject Island
LBC Violation
Relative Clause Island
Sentential Subject Island
CNPC Island
Whether Island

The only manipulation between the two was in choice of reference sentence and incidental sample size.

Results

As before, the responses were divided by the reference judgment and log-transformed prior to analysis. First, repeated measures linear regressions were performed on the means for each experiment. Summary tables of the y-intercept, slope, and p-value are included below. Significant effects, are marked in **bold**:

Table 4.18. Linear regressions for means, CSC-reference

	(b)	Exp(b)	p value
Adjunct Island	-0.13	-0.001	.97
CNPC Island	-0.03	0.005	.73
CSC violation	0.02	-0.004	.69
Infinitival Sentential Subject Island	-0.35	0.020	.07
LBC violation	-0.10	-0.003	.84
Relative Clause Island	-0.09	-0.005	.78
Sentential Subject Island	-0.19	0.030	.14
Whether Island	-0.03	-0.004	.79

Repeated measures linear regressions were also performed on the residuals:

Table 4.19. Linear regressions for means, If-reference

	(b)	Exp(b)	p value
Adjunct Island	-0.24	0.01	.64
CNPC Island	-0.14	-0.01	.64
CSC violation	0.38	0.03	.18
Infinitival Sentential Subject Island	**-0.49**	**0.05**	**.003**
LBC violation	-0.18	-0.01	.43
Relative Clause Island	-0.29	0.02	.23
Sentential Subject Island	-0.40	0.02	.28
Whether Island	**-0.16**	**0.02**	**.05**

Table 4.20. Linear regressions for residuals, CSC-reference

	(b)	Exp(b)	p value
Adjunct Island	0.23	0.00	.63
CNPC Island	0.18	0.01	.61
CSC violation	0.04	0.01	.12
Infinitival Sentential Subject Island	0.32	0.00	.93
LBC violation	**0.17**	**0.02**	**.02**
Relative Clause Island	**0.15**	**0.03**	**.02**
Sentential Subject Island	0.20	0.01	.72
Whether Island	**0.11**	**0.03**	**.02**

Discussion

The effects appear to break down like this. First, there were three significant effects with the CSC-reference, all of which appeared in the variance (or spread) of the judgments: Left Branch Constraint violations, Relative Clause Islands, and Whether Islands. However, these three effects were not replicated with the If-reference, as there were no significant effects of variance. There were two significant effects in means with the If-reference: the Infinitival Sentential Subject Islands and Whether Islands. Despite the effects showing up in two different measures (variance versus mean), this does appear to be a partial replication (with respect to Whether Islands) between these two experiments. The question, then, is what to make of it.

Table 4.21. Linear regressions for residuals, If-reference

	(b)	Exp(b)	p value
Adjunct Island	0.26	-0.01	.45
CNPC Island	0.18	0.01	.30
CSC violation	0.40	-0.03	.12
Infinitival Sentential Subject Island	0.32	-0.01	.32
LBC violation	0.30	-0.01	.23
Relative Clause Island	0.27	0.00	.81
Sentential Subject Island	0.32	0.00	.90
Whether Island	0.17	0.00	.99

Unfortunately, the answer seems to be that not too much can be made of it. First, it should be noted that there were 32 statistical analyses conducted in this analysis with direct comparisons of at least 4 conditions at a time (the means and variances of each island across the two experiments), and upwards of 16 or 32 comparisons. Given the nature of probabilities, the more analyses one performs, the more likely a significant result will be. In fact, with a target p-value of .05, 20 analyses will nearly guarantee a significant result. One of the most conservative corrections for this problem is the Bonferroni correction. The Bonferroni corrected p-value for 4 comparisons is .0125. The only significant effect that achieves this level is in the mean of Infinitival Sentential Subject Islands with the If-reference. None of the results reach significance when corrections are made for larger numbers of comparisons such as 16 or 32.

As one anonymous reviewer points out, there may be more going on in the If-reference experiment given that the If Island used as a reference is structurally similar to Whether Islands. For instance, it could be the case that this structural equivalence means that the participants are in fact seeing 55 instances of Whether

Islands in this experiment, and that the extreme number of repetitions is what causes the significant increase in mean judgment for Whether Islands. Setting aside the previous argument that the correct p-value for Whether Islands is not significant, there are other reasons that this argument does not go through. First, under this conception BOTH the reference If Island AND the experimental Whether Islands should be affected by satiation. If that were true, there should be no effect on the Whether Islands at all, since both the reference and the experiment items would be increasing in acceptability together. The fact that there is an effect (for the sake of argument; after Bonferroni correction there is no effect) indicates that the reference If Islands and the Whether Islands were being treated differently by the participants. Furthermore, if the reference sentence were indeed increasing in acceptability, we would expect to find *negative* satiation, that is, decreases in acceptability, for the other violations (unless, of course, they were satiating at the same rate as the reference). Since there were no significant negative effects, it does not seem like the If-reference was satiating at all.

So the answer to whether judgments are stable given the magnitude estimation task and unbalanced designs is a guarded yes. There was one true instability effect, but much like the instability found in the yes/no task, it is not overwhelming evidence for a satiation effect.

4.5 Conclusion

Now we are in a position to fill in all four cells of our crossed design:

Table 4.22. Crossed design of factors Task and Design

	Yes/No	MagE
Balanced	stable	stable
Unbalanced	unstable	stable

What we've found is that acceptability judgments are strikingly stable within balanced designs. There is instability with unbalanced designs and yes/no tasks, although it seems that magnitude estimation tasks are more resilient to the effect of unbalanced designs. This suggests a standard interaction effect: unbalanced designs leads to instability, but more instability for yes/no tasks than magnitude estimation tasks. The replication problem for previous satiation studies now receives a natural explanation: violation type is not the major factor determining instability, the interaction of task and design are, perhaps due to a response strategy in which participants (consciously or unconsciously) attempt to balance the number of yes and no responses over the course of the experiment.

The implications for syntactic theory and linguistic methodology are straightforward. Snyder's ingenious response to claims that satiation undermines syntactic theory can no longer hold, but it does not have to. Given that satiation is most likely an artifact of design choices, it no longer threatens to undermine acceptability judgments as a robust form of data. Of course, if satiation is an artifact, it can no longer serve as a source of data for syntacticians to discriminate between different types of grammatical knowledge, or as data for determining the natural syntactic classes of violations. But that is a small price to pay for the empirical benefit of data that is stable over repeated measurements.

4.6 Some remaining questions

The conclusion that instability can be avoided through balanced designs is not to say that there are not questions about the judgment task that are ripe for future research. As a final section, I review two such questions, and propose starting points for the investigation.

4.6.1 Whether Islands versus the That-trace effect

Statistical tests aside, one does get the impression that there is a pattern throughout the experiments presented in this chapter: Whether Islands arise in the discussion of instability more often than any other violation, and some violations, for instance That-trace effects, never arise. Even given the analysis of instability presented in this paper, it still may be the case that only certain violations can be unstable (and conversely that some are always stable). To be clear, given that the instability has to be licensed by very specific design factors, this is not saying that there may be a new classification system. The claim would have to be the weaker claim that some violations are susceptible to instability (not unstable by definition), and others are not. This suggests that susceptibility to instability may be a side-effect of the judgment process itself, and how it interacts with the nature of certain violations.

For instance, the fact that That-trace effects are never susceptible to instability could reduce to the fact that That-trace effects are *correctable*, as demonstrated experimentally by Crain and Fodor 1987 in their discussion of the sentence matching

task. Because participants can easily identify the source of the violation, their judgments may become 'anchored' in a way that is not possible with structural violations that cannot be easily corrected, such as Whether Islands. And the fact that Whether Islands seem susceptible to instability while Sentential Subject Islands do not may then reduce to relative acceptability: non-correctable violations that are closer to the yes/no threshold would logically be more likely to cross that threshold. Interestingly, Snyder rejects relative acceptability as an explanation for the satiating versus non-satiating violations based on the fact that (in a scale-based rating study with 10 participants and no error terms reported), the order of relative acceptability from highest to lowest was:

Table 4.23. Relative acceptability versus satiation from Snyder 2000

Relative acceptability	**Satiating violations**
Want-for violations	Whether Islands
Whether Islands	CNPC Islands
That-trace effects	Subject Islands Subject Islands
CNPC Islands	
Adjunct Islands	
LBC violations	

As laid out, there is no direct relationship between relative acceptability and satiation. However, by using the correctable/non-correctable classification, and removing Want-for and That-trace from the paradigm since they are both correctable by the removal of a single word (*for* or *that*), the relatively acceptability order corresponds almost directly with the satiation results:

Table 4.24. Relative acceptability versus satiation based on non-correctability

Relative acceptability	**Satiating violations**
Whether Islands	Whether Islands
Subject Islands	CNPC Islands
CNPC Islands	Subject Islands
Adjunct Islands	
LBC violations	

In fact, there is at least anecdotal independent evidence for such an account: during the debriefing of participants after the Snyder replications reported in section 2, six participants reported noticing a difference among the unacceptable sentences in that some of them could be corrected by "changing a *for* or *that*". Given the results of Crain and Fodor 1987 and their potential relevance for understanding the complete picture of judgment instability, there is obviously room for future research into the effect of non-syntactic factors such as correctability on acceptability judgments.

4.6.2 What about the models of satiation?

Because of the replication problem of satiation, the experiments in this paper can only go so far toward identifying the nature of the instability that gives rise to satiation effects as defined in Snyder 2000. What little we do know is this: magnitude estimation studies are resilient to instability such that there is no strong evidence for changes in mean or variance, that is, there is no evidence for model 1 or model 2. Unfortunately, it is not clear whether this is because these models do not capture the type of instability seen in yes/no tasks or because magnitude estimation tasks are too stable. Compounding this problem is the fact that we do not yet have validated

methodologies for investigating model 3, a change in the yes/no threshold itself, given the complexity of the judgment involved in such a categorical distinction.

As briefly mentioned at the beginning of the chapter, one of the of replications was run with an additional task: participants were asked to rate their confidence in their yes/no response on a 7-point scale following each judgment. The idea was that there might be a correlation between violations that satiate and violations that lead to lower confidence in judgments over time. For instance, it is plausible that a changing category threshold may lead to decreasing confidence in judgments, thus confidence could track threshold instability (although there are many other reasons for confidence to change over time). Unfortunately, as we have seen, there were no satiation effects by Snyder's definition in this experiment, so it is impossible to draw the intended correlations. However, despite the lack of satiation in this experiment, there were significant decreases in confidence for Adjunct Islands, CNPC Islands, That-trace effects, and Want-for violations. Without corresponding satiation effects it is hard to draw conclusions from these decreases in confidence. However, it is suggestive that future research on the factors influencing participants' confidence about their judgments could be correlated with the factors that we have seen influence stability, namely choice of task and choice of design.

Chapter 5

Processing effects and the differential sensitivity of acceptability

Linguists have agreed since at least Chomsky 1965 that acceptability judgments are too coarse grained to distinguish between effects of grammatical knowledge (what Chomsky 1965 would call competence effects) and effects of implementing that knowledge (or performance effects). With the rise of experimental methodologies for collecting acceptability judgments, there has been a renewed interest in attempting to identify the contribution of performance factors, in particular processing factors, to acceptability judgments. For instance, Fanselow and Frisch 2004 report that local ambiguity in German can lead to increases in acceptability, suggesting that the momentary possibility of two representation can affect acceptability. Sag et al. submitted report that factors affecting the acceptability of Superiority violations also affect the processing of wh-questions as measured in reading times, suggesting that there might be a correlation between processing factors and the acceptability of Superiority violations. This chapter builds on this work by investigating three different types of temporary representations created by the Active Filling processing strategy (Frazier and Flores d'Arcais 1989) to determine if they affect the acceptability of the final representation. The question is whether every type of processing effect that arises due to the active filling strategy affects acceptability, or whether acceptability is differentially sensitive to such processing effects. The results suggest that judgment tasks are

indeed differentially sensitive: they are sensitive to some processing effects, but not others. This differential sensitivity in turn suggests that further research is required to determine the class of processing effects that affect acceptability in order to refine the relationship between acceptability, grammaticality, and processing effects. Determining that relationship will be the first step toward assessing the merits of both processing-based and grammatical knowledge-based explanations of acceptability.

5.1 Syntactically and semantically ungrammatical representations

5.1.1 Some psycholinguistic background

The experiments in this chapter build upon one of the major findings of sentence processing research: the active filling strategy. The active filling strategy is defined by Frazier and Flores d'Arcais (1989) as when a filler has been identified, rank the possibility of assigning it to a gap above all other options. Or, in other words, the human parser prefers to complete long distance dependencies as quickly as possible. Because the quickest possible completion site is not always the correct one, the active filling strategy entails the construction of many temporary, incorrect representations. The experiments in this chapter take advantage of these temporary representations.

One of the major pieces of evidence for the active filling strategy is the filled-gap effect. Simply put, the filled-gap effect arises when the parser completes a wh-dependency with a verb that subsequently turns out to have an object, and thus no free thematic positions. Because sentence processing in English proceeds from left to

right, potentially transitive verbs are encountered prior to their objects. The active filling strategy mandates that the parser complete an open wh-dependency at the first appropriate verb. If after the dependency is completed, the parser encounters an object of the verb, there is a corresponding slow-down in reading times (at the object of the verb) due to the competition of the two NPs for the object thematic position of the verb. Stowe 1986 demonstrated this effect with the following quadruplet:

(26) My brother wanted to know...

a. if Ruth will bring us home to Mom at Christmas.

b. who$_1$ t$_1$ will bring us home to Mom at Christmas.

c. who$_1$ Ruth will bring t$_1$ home to Mom at Christmas.

d. who$_1$ Ruth will bring us home to t$_1$ at Christmas.

Stowe found a significant reading time slow-down at the position of the object *us* when there is a displaced wh-filler (26d) as compared to when there is no displaced wh-filler (26a):

Table 5.1. Reading times (in ms) at critical words for each dependency, from Stowe 1986

	Ruth	*us*	*Mom*
None (if)	661	**755**	755
WH-Subject	—	801	812
WH-Object	680	—	833
WH-Preposition	689	**970**	—

The plausibility effect is a second piece of evidence for the active filling strategy. The plausibility effect is a slow-down in reading times caused when the completed dependency between a wh-filler and a verb is semantically implausible. The active

filling strategy mandates that the dependency be completed as soon as possible, regardless of the ensuing semantic anomaly. As such, there is a reading time slow-down after the verb when the semantic anomaly is detected. For instance, Pickering and Traxler (2003) found a significant slow-down at the verb when the displaced wh-filler is an implausible object of the verb killed:

(27) a. That's the **general**$_1$ that the soldier **killed** enthusiastically for t_1 during the war in Korea.

b. % That's the **country** that the soldier **killed** enthusiastically for t_1 during the war in Korea.

Table 5.2. Reading times at critical words by filler type, from Pickering and Traxler 2003

	killed enthusiastically
Plausible	1045 ms
Implausible	1157 ms

5.1.2 Rationale

Given that the question under investigation is whether temporary representations have an effect on the judgment of the final representation, it goes without saying that it must be established that the temporary representations being manipulated are actually constructed. The filled-gap effect (Crain and Fodor 1985, Stowe 1986) and the plausibility effect (Garnesey et al. 1989, Tanenhaus et al. 1989) were chosen because their effects are so well-established that they serve as tools for investigating filler-gap dependencies in the processing literature. Furthermore, despite both being

reflexes of active filling, the source of the effect for each paradigm is different. The slow-down that occurs in the filled-gap paradigm is due to an argument structure violation, resulting in a temporary structure in which there is an argument structure violation. On the other hand, the slow-down in the plausibility paradigm is due to a temporary representation with a semantically anomalous interpretation. Thus these two paradigms are ideal for investigating the effects of different types of temporarily illicit representations.

Because all of the target conditions must ultimately be acceptable sentences to avoid any interfering grammaticality effects, three design elements were incorporated into experiment 1 to ensure that a failure to detect either the filled-gap or the plausibility effect was not due to a lack of sensitivity. First, the task chosen was magnitude estimation. Recent findings have indicated that magnitude estimation is well suited for detecting differences among acceptable sentences (e.g., Featherston 2005b). Second, a relatively large number of participants were asked to participate in this study. Recent work has suggested that reliable results can be obtained from samples as small as 10 (Myers 2006), therefore the large sample size for experiment 1 (N=86) should be adequate to detect even very small differences. Finally, a third condition set was included to determine whether the task and participant pool were sensitive to distinctions among acceptable sentences. The third condition set was taken from an ERP study of the distance between a wh-filler and its gap by Phillips et al. (2005). Phillips et al. manipulated the distance by displacing the wh-filler either 1 or 2 clauses away from the gap position:

(28) a. The detective hoped [that the lieutenant knew [**which accomplice** the shrewd witness would recognize __ in the lineup]].

b. The lieutenant knew [**which accomplice** the detective hoped [that the shrewd witness would recognize __ in the lineup]].

Phillips et al. found a delay in the onset of the P600, a brain response that has been linked to the association of a wh-filler with its gap, for the Long WH condition, which they interpret as a reflex of the time it takes to retrieve the stored filler from working memory (longer distance = longer retrieval time, perhaps because of a decaying representation). But crucial to our purposes, Phillips et al. conducted a ratings survey in which they asked the participants to rate the complexity of the two conditions on a scale from 1 to 5:

Table 5.3. Mean complexity ratings for short and long movement, from Phillips et al. 2005

	Mean	Standard Deviation
Short – 1 clause	2.71	0.65
Long – 2 clauses	3.51	0.51

The difference between the two conditions was highly significant ($t(23)=5.83$, $p<.001$), indicating that judgment tasks could indeed detect an effect that leads to unconscious processing effects. Thus this condition was included in experiment 1 as a baseline to test the sensitivity of the task and participant pool (although experiment 1 was an acceptability rating task rather than complexity rating task).

Participants

86 University of Maryland undergraduates participated for extra credit. All of the participants were self-reported native speakers of English. The survey was 36 items long including practice items, and took about 15 minutes to complete.

Materials and Design

The design included 3 condition sets: the filled-gap paradigm to test temporary syntactically ungrammatical representation, the plausibility paradigm to test temporary semantically ungrammatical representations, and the wh-distance paradigm to check the sensitivity of the task and participant pool. For the filled-gap condition set of this experiment, the WH-Object and WH-Preposition conditions from Stowe 1986 were reconstructed:

(29) The Filled-gap effect condition set: Gap and Filled-gap

a. My brother wanted to know **who** Ruth will bring __ home to Mom at Christmas.

b. My brother wanted to know **who** Ruther will bring **us** home to __ at Christmas.

In the filled-gap condition (29b) , the failure to integrate the displaced wh-filler with the verb creates a representation in which the dependency is incomplete, which persists until the gap in the prepositional phrase. The materials for the plausibility condition set were taken directly from the published materials of Pickering and Traxler 2003, although an additional adverb was added to each token to increase the duration of the semantically ungrammatical representation, and the matrix clause was changed

to match the style of the filled-gap conditions (i.e., declarative sentences):

(30) The Plausibility effect condition set: Plausible and Implausible

a. John wondered **which general** the soldier killed __ effectively and enthusiastically for __ during the war in Korea.

b. John wondered **which country** the soldier killed __ effectively and enthusiastically for __ during the war in Korea.

The implausible condition (30b) creates a dependency at the verb killed that is semantically ungrammatical, which persists through the two adverbs until the gap in the prepositional phrase. The materials for the wh-distance condition set were taken directly from the published materials of Phillips et al. 2005:

(31) WH-movement distance condition set: Short and Long

a. The detective hoped [that the lieutenant knew [**which accomplice** the shrewd witness would recognize __ in the lineup]].

b. The lieutenant knew [**which accomplice** the detective hoped [that the shrewd witness would recognize __ in the lineup]].

There were 6 total conditions (2 each of 3 sets) under investigation. 24 lexicalizations of the filled-gap conditions were reconstructed following the examples from Stowe 1986. 24 lexicalizations of the wh-distance conditions were taken from the materials of Phillips et al. 2005. Only 12 lexicalizations were available for the plausibility conditions from Pickering and Traxler 2003. A 24-cell Latin Square was constructed such that each list contained 2 tokens of each condition. 14 unacceptable fillers (various syntactic island violations) were added, and each list was pseudo-randomized

such that no more than 2 target conditions were consecutive, and no related conditions were consecutive. 8 practice items were added, resulting in a 34 item survey. The instructions were a modified version of the instructions distributed with the WebExp software suite (Keller et al. 1998). The reference sentence for both the practice and experimental items was a three-clause long sentence containing a whether-island violation: *Mary figured out what her mother wondered whether she was hiding.*

Results

Results were divided by the reference score and log transformed prior to analysis. Paired t-tests were performed on each of the condition pairs:

Table 5.4. Results and paired t-tests for experiment 1

	mean	SD	df	t	*p*	r
long-distance	.08	.19				
short-distance	.20	.17	85	5.324	.001	.50
filled-gap	.03	.24				
unfilled-gap	.16	.24	85	5.616	.001	.52
implausible	.09	.18				
plausible	.10	.20	85	0.514	.608	—

As the chart indicates, there was a large and highly significant decrease in acceptability for longer wh-dependencies, and in exactly the same direction as obtained by Phillips et al 2005. There was also a large and highly significant decrease in acceptability for filled-gaps, mirroring the direction of the effect found by Stowe 1986. However, there was no effect of plausibility. Even though there are no direct statistical comparisons across the groups, it is clear that both of the significant p values are well under the conservative Bonferroni correction level of .0167.

Discussion

The pattern of results from experiment 1 is extremely puzzling. First, it is clear from the wh-distance effect that the design is capable of detecting differences that lead to processing effects. However, when it comes to the two active filling effects, a significant effect was only found for the filled-gap effect. Also, given the large sample size, it seems unlikely that increasing the sample size will lead to an effect of plausibility. At first glance, this seems to suggest that temporary syntactic ungrammaticality affects global judgments, whereas temporary semantic ungrammaticality does not. Unfortunately, there is a second possible explanation: reanalysis. By definition, the filled-gap condition of the filled-gap paradigm involves abandoning one structure and constructing a second structure, a type of syntactic reanalysis: when the association between the wh-filler and the thematically saturated verb fails, the parser must reanalyze the structure such that the wh-filler is then associated with the preposition. In other words, the parser attempts to 'drop the filler twice. However, the true gap condition of the paradigm involves no such reanalysis because the first association with the verb succeeds. It could be the case then that the difference in acceptability between the two conditions is an effect of reanalysis on the judgment. This would also account for the lack of effect in the plausibility conditions: in both conditions, the wh-filler is initially associated with the verb and later reanalyzed as the object of the preposition. Thus if reanalysis leads to a decrease in acceptability, one would expect an effect in the filled-gap paradigm but no in the plausibility paradigm. Experiment 2 was designed tease apart these two hypotheses (asymmetry due to temporary unacceptability versus asymmetry due to reanalysis).

5.2 The reanalysis confound

Unfortunately, by definition there is no way to eliminate reanalysis from the filled-gap and plausibility paradigms. However, it is possible to add reanalysis to the true gap condition of the filled-gap paradigm, thus making it completely parallel to the plausibility paradigm in that both conditions will contain reanalysis. If the asymmetry in the presence of reanalysis across the two paradigms was the source of the asymmetry in the results for experiment 1, then eliminating the reanalysis asymmetry should eliminate the asymmetry in the results such that both paradigms return no effect. Experiment 2 was designed to test this hypothesis. Furthermore, by adding the true-gap condition from experiment 1 that lacks reanalysis and comparing it the new true-gap + reanalysis condition, it is possible to isolate the effect of reanalysis alone, if it should exist. This comparison investigates the effect of a temporary grammatical representation on the judgment of the final representation, or in other words, the effect of processing difficulty without ungrammaticality, setting up the three-way comparison of temporary representations discussed in section 1.

Participants

21 University of Maryland undergraduates participated in this experiment. All were self-reported native speakers of English without any formal training in linguistics. All were paid for their participation.

Materials and Design

The materials for experiment 2 were adapted from the materials for the plausibility conditions in experiment 1, which were themselves adapted from the published

materials of Pickering and Traxler 2003. These materials were chosen for two reasons: (i) the plausibility materials already contained the necessary structure to include reanalysis in both the filled-gap and true-gap conditions; and (ii) if a filled-gap effect is indeed found using these materials, it would serve to exclude the possibility that the lack of effect for plausibility in experiment 1 was due to the meanings of the materials. Three conditions were used to test whether the source of the asymmetry from experiment 1 was the reanalysis asymmetry. First, a filled-gap condition was constructed out of the materials from Pickering and Traxler 2003. Next a true-gap condition was constructed with an additional gap in the prepositional phrase to represent a gap+reanalysis condition. Finally, a standard true-gap condition was constructed with no gap in the prepositional phrase to serve as both a replication of the filled-gap effect in experiment 1, and to serve as a control for gap+reanalysis condition so that an effect of reanalysis alone could be tested:

(32) *Filled-gap + reanalysis (FG+R)*

John wondered **which general** the soldier killed **the enemy** effectively and enthusiastically for __ during the war in Korea.

Gap + reanalysis (G+R)

John wondered **which general** the soldier killed __ effectively and enthusiastically for __ during the war in Korea.

Gap (G)

John wondered **which general** the soldier killed __ effectively and enthusiastically for our side during the war in Korea.

Again, the competing hypotheses make different predictions: if reanalysis is the source of the asymmetry, then experiment 2 should yield no effect between FG+R and G+R because both conditions involve reanalysis, and a significant effect between G+R and G since there is an asymmetry in reanalysis; if the asymmetry is due to the nature of the representation constructed, then there should again be an effect between FG+R and G and also an effect between FG+R and G+R. This hypothesis makes no prediction about G+R and G, but that comparison would indicate whether reanalysis has any effect at all. 8 lexicalizations of each triplet were constructed and distributed using a Latin Square design. Each list contained 1 token of each condition. 10 additional conditions from an unrelated study were included as fillers. By hypothesis these 4 of these fillers were considered acceptable, while 6 were considered unacceptable, yielding a nearly balanced ratio of acceptable/unacceptable. 8 practice items were included for a total of 21 items. The task was magnitude estimation, and the instructions were identical to those of experiment 1. The reference sentence was also identical.

Results

As before, results were divided by the reference score and log-transformed prior to analysis:

Table 5.5. Results for experiment 2

	mean	SD
filled-gap	-.02	.22
gap + reanalysis	.09	.22
gap only	.11	.20

There was a large and significant effect of FG+R versus G+R ($t(20)=2.8$,

p=.005, r=.53), and as expected of FG+R versus G (t(20)=2.8, p=.005, r=.53). There was no effect of G+R versus G (t(20)=0.32, p=.37). And although all of the p values were one-tailed, it should be noted that both of the significant p values were well below the Bonferroni corrected level of .017, even at their two-tailed value of p=.01.

Discussion

By introducing a second gap within the prepositional phrase of the gap condition, experiment 2 was able to eliminate the asymmetry of reanalysis from the design of the filled-gap paradigm, and thus tease apart the two possible explanations of the asymmetry in the results of experiment 1. The persistence of the effect despite the introduction of reanalysis into both conditions confirms that there is something peculiar to the filled-gap effect that affects the judgment of the final representation. Furthermore, the lack of effect between the two gap conditions suggests that reanalysis has no lasting effect on the judgment of the final representation: apparently there is no lasting cost associated with abandoning one well-formed representation for another.[1]

[1]Because there was no comprehension task included in experiment 2, it is possible that the lack of effect of reanalysis actually represents a lack of reanalysis, in that the participants might not notice the gap position in the string *for during*. Of course, if it was the case that *for during* was not an appropriate cue for a gap, then it would be unclear why there was no effect of plausibility in experiment 1, as without reanalysis the implausible condition is actually unacceptable, and should have received a correspondingly low judgment.

5.3 The differential sensitivity of acceptability to processing effects

At an empirical level, the results from the experiments in this chapter reveal a surprising asymmetry in the effects of temporary representations on global acceptability, suggesting that syntactic difficulties are treated by the judgment process in a qualitatively different way than semantic or processing difficulties. This seems to indicate that judgment tasks are tapping directly into syntactic knowledge in a very real sense. At a methodological level, these results demonstrate the sensitivity of formal judgment experiments: the ability to detect significant differences between two acceptable sentences opens the possibility of using judgment experiments to explore phenomena that are typically the domain of sentence processing studies. And at a theoretical level, these results indicate that some, but not all, processing effects affect acceptability judgments. This differential sensitivity suggests that it is possible to use acceptability judgments to investigate the predictions of processing-based analyses of acceptability facts by first determining whether the processing effects in question affect acceptability at all, and then whether the acceptability of theoretically related phenomena are similarly affected (or unaffected).

Chapter 6

The role of acceptability in theories of wh-in-situ

One of the most salient properties of human language is the presence of non-local dependencies. One of the major goals of syntactic theory over the past 40 years has been to classify the properties of these dependencies, and ultimately attempt to explain them with the fewest number of dependency constructing operations. Yet even when attention is restricted to wh-questions, there seem to be several different types of dependencies, each requiring slightly different operations. As such, it is not surprising that there are a number of different proposals in the literature to capture wh-dependencies: feature movement, phrasal movement, choice-function application, long distance AGREE, et cetera. What is surprising is that in many ways the field of syntax has decided that acceptability judgments can provide little additional insight into the nature of these dependencies. One of the major factors contributing to this situation is the fact that the structural restrictions on wh-dependencies seem to be fairly straightforward: overt wh-displacement leads to Island effects, wh-in-situ (of arguments) is not constrained by Islands. The Island facts thus form the basis from which all analyses must begin. The data that constitutes evidence for or against these analyses usually comes from either i) non-wh dependencies that also use the postulated operation, or ii) the nature of the possible answers to the different types of wh-questions (see especially Dayal 2006 for a review).

While previous chapters in this dissertation focused on the relationship between acceptability and grammaticallity in an attempt to refine our understanding of the nature of grammatical knowledge, this chapter demonstrates a more direct relationship between understanding the source of acceptability and syntactic theories. The claim in this chapter is straightforward: there are new types of acceptability data that can be revealed through experimentation, and a better understanding of the relationship between these new acceptability effects and grammatical knowledge can have significant consequences for the set of possible dependency forming operations. This chapter focuses almost exclusively on wh-in-situ in multiple wh-questions in English. In-situ wh-words must be interpreted somehow, and that interpretation is dependent on the interpretation of the wh-word in matrix C. There is a general consensus that there must be a dependency between these two positions, but there is significant debate over the nature of that dependency, and in particular, over the dependency forming operation(s) that create it. Various proposals have been made in the literature such as covert wh-movement (Huang 1982), null operator movement and unselective binding (Tsai 1994), choice-function application and existential closure (Reinhart 1997), overt movement and pronunciation of the lower copy (Bošković 2002), and long distance AGREE (Chomsky 2000).

Section 1 provides the first discussion of new data, focusing on Huang's (1982) claim that there are no Island effects with wh-in-situ in English. A series of experiments are presented that demonstrate the existence of Subject Island effects with wh-in-situ, but no other Island effects. Section 2 is the second data section, presenting evidence that the distance between a wh-in-situ and the higher wh-phrase

affects acceptability, while the distance of similarly complex long distance dependencies, such as binding dependencies do not. Section 3 discusses the consequence of these new data points on the various proposals for wh-in-situ dependencies in English. The general conclusion is that these facts suggest a movement-based account such as overt movement with lower copy pronunciation, covert movement, or null operator movement, and raise serious difficulties for non-movement approaches such as choice-function application and long distance AGREE.

6.1 Island effects and wh-in-situ in English

Working under the assumption that the dependency between an in-situ wh-word and the matrix [+wh] C is formed through covert movement, Huang 1982 argues that there is a direct parallelism between wh-in-situ languages like Chinese and the wh-in-situ that occurs in multiple wh-questions in English: both undergo covert movement, but (in the case of wh-arguments) neither show Island effects. This can be seen most directly in English with a Whether Island:

(33) a. * What do you wonder whether John bought?

b. Who wonders whether John bought what?

The first example is just a standard Whether Island effect with overt wh-movement. In the second example, the in-situ wh-word has matrix scope (because it must be answered), indicating a dependency with the matrix C, but there is no (apparent) ensuing Whether Island effect. Huang argues that this is also the case for CNPC Islands:

(34) a. * What did you make the claim that John bought?

b. Who made the claim that John bought what?

Huang used these facts together with the lack of Island effects in wh-in-situ (of wh-arguments) in Chinese to argue that covert movement is not constrained by Subjacency.

Insofar as overt displacement and constraint by Subjacency are characteristic properties of movement, it is easy to see how the lack of Island effects with wh-in-situ is a principal component of analyses in which the dependency between the matrix C and wh-in-situ is accomplished with mechanisms other than standard wh-movement. Therefore, the first step of this study was to evaluate Huang's claim that there are no Island effects with wh-in-situ in English with the major Island types. As will become clear momentarily, while Huang's claim is mostly correct, the comparisons such as the ones above actually obscure potential evidence for covert movement Island effects.

6.1.1 Testing Huang 1982

Participants

92 University of Maryland undergraduates participated in this experiment for extra course credit. All were self-reported native speakers of English.

Materials and Design

6 Island types were tested following the paired contrasts for CNPC and Whether Islands published in Huang 1982. The (a) condition in each pair is an Island violation with overt wh-movement. The (b) condition in each pair is a multiple wh-question in

which the in-situ wh-word covertly moves out of an Island structure.

(35) Adjunct Island

a. What does Mary study music because her boyfriend plays?

b. Who studies music because her boyfriend plays what?

(36) CNPC Island

a. What did you doubt the claim that Jesse invented?

b. Who doubted the claim that Jesse invented what?

(37) CSC violation

a. Who did you claim the bully teased his brother and ?

b. Who claimed the bully teased his brother and who?

(38) Relative Clause Island

a. What did Chris slap the man that stole?

b. Who slapped the man that stole what?

(39) Subject Island

a. What does John think that a story about would be scary?

b. Who thinks that a story about what would be scary?

(40) Whether Island

a. What did you wonder whether the detective found?

b. Who wonders whether the detective found what?

12 lexicalizations were created for each condition. Lexicalizations were controlled for length in number of words within each condition set. These 12 lexicalizations were combined with 10 conditions, each with 12 lexicalizations, from an unrelated experiment, and distributed using a Latin Square design for a total of 12 lists. The 12 lists were pseudorandomized such that related conditions were never consecutive.

The task was magnitude estimation. The instructions were a modified version of the instructions published with the WebExp software suite (Keller et al. 1998). The reference sentence was: *What did Mary wonder whether her mother was buying for her father?*

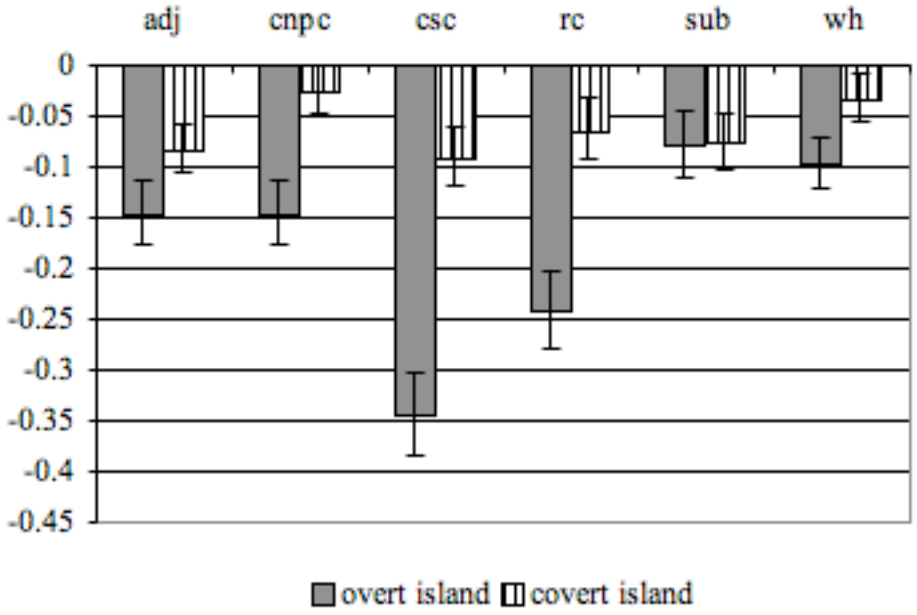
Results

Responses were divided by the value of the reference sentence and log-transformed prior to graphing and analysis. Paired t-tests were performed on each Island type (one-tailed). The results are summarized in the following table and chart, in which it is clear that covert movement out of an Island is significantly more acceptable than overt movement out of an Island for every Island type except Subject Islands. There is no significant difference between overt and covert movement out of Subject Islands:

Table 6.1. Results and paired t-tests for Huang-style conditions

	overt movement		covert movement					
	mean	SD	mean	SD	df	t	p	r
Adjunct	-.15	.31	-.08	.23	91	1.73	.044	.18
CNPC	-.15	.30	-.02	.23	91	3.16	.001	.31
CSC	-.34	.39	-.09	.28	91	5.52	.001	.50
Rel. Cl.	-.24	.37	-.06	.30	91	5.10	.001	.47
Subject	-.08	.31	-.08	.26	91	0.09	.465	—
Whether	-.10	.24	-.03	.23	91	2.15	.017	.22

Figure 6.1. Overt versus covert Island effects following Huang 1982



Discussion

Unsurprisingly, the results of this experiment confirm the claims from Huang 1982 for CNPC and Whether Islands, and extend to Adjunct, CSC, and Relative Clause Islands. However, there is no difference between overt and covert movement for Subject Islands, suggesting that Subject Islands are different than the other Island types. Unfortunately, there is no way to directly interpret the lack of effect with Subject Islands: it could be that there is a covert movement Island effect, or that there is no overt movement Island effect, or even that overt Subject Islands are weak enough, and the additional complexity of multiple wh-questions strong enough, to lead to similar acceptability. In fact, because Island effects were not defined across conditions (the Island structures were tested without non-Island controls) we cannot directly interpret the effects that were found for the other Island types: the strongest

claim that can be made is that the overt movement Island effect is stronger than the covert movement Island effect; we cannot actually claim that there are no covert movement Island effects. To get around these confounds, a second experiment was run using the designs in which Island effects are defined as interactions of two factors: STRUCTURE and MOVEMENT.

6.1.2 A better design for investigating covert movement Islands

In this experiment, wh-in-situ Island effects are defined as an interaction: the difference in acceptability between wh-in-situ in an Island structure and wh-in-situ in a minimally different non-Island structure is compared to the difference in acceptability between the two structures themselves without any wh-in-situ. In other words, there are two factors STRUCTURE and WH-IN-SITU each with two levels. So for each Island type, there are four conditions:

(41) Adjunct Island

  i. Complement

    i. Who$_1$ t$_1$ suspects [$_{CP}$ that you left the keys in the car?]

    ii. Who$_1$ t$_1$ suspects [$_{CP}$ that you left what in the car?]

  ii. Adjunct

    i. Who$_1$ t$_1$ worries [$_{ADJ}$ that you leave the keys in the car?]

    ii. Who$_1$ t$_1$ worries [$_{ADJ}$ that you leave what in the car?]

(42) CNPC Island

i. CP Complement

i. Who$_1$ t$_1$ denied [$_{CP}$ that you could afford the house?]

ii. Who$_1$ t$_1$ denied [$_{CP}$ that you could afford what?]

ii. NP Complement

i. Who$_1$ t$_1$ denied [$_{NP}$ the fact that you could afford the house?]

ii. Who$_1$ t$_1$ denied [$_{NP}$ the fact that you could afford what?]

(43) Subject Island

i. Simple NPs

i. Who$_1$ t$_1$ thinks the speech interrupted what?

ii. Who$_1$ t$_1$ thinks what interrupted the TV show?

ii. Complex NPs

i. Who$_1$ t$_1$ thinks [the speech by the president] interrupted [the TV show about what]?

ii. * Who$_1$t$_1$ thinks [the speech by who] interrupted [the TV show about whales]?

(44) Whether Island

i. CP Complement

i. Who$_1$ t$_1$ thinks [$_{CP}$ that you wrote the letter?]

ii. Who$_1$ t$_1$ thinks [$_{CP}$ that you wrote what?]

ii. Whether Complement

i. Who$_1$ t$_1$ wonders [$_Q$ whether you wrote the letter?]

ii. Who$_1$ t$_1$ wonders [$_Q$ whether you wrote what?]

(45) Specificity Island

i. Non-specific

i. Who$_1$ t$_1$ thinks that you read [a book about Egypt]?

ii. Who$_1$ t$_1$ thinks that you read [a book about what]?

ii. Specific

i. Who$_1$ t$_1$ thinks that you read [John's book about Egypt]?

ii. Who$_1$ t$_1$ thinks that you read [John's book about what]?

(46) Relative Clause Island

i. CP Complement

i. Who$_1$ t$_1$ knows [$_{CP}$ that the woman read a book?

ii. Who$_1$ t$_1$ knows [$_{CP}$ that the woman read what?

ii. NP Complement

i. Who$_1$ t$_1$ knows [$_{NP}$ the woman that read a book?]

ii. Who$_1$ t$_1$ knows [$_{NP}$ the woman that read what?]

Specificity Islands were included in this follow-up study for comparison to Subject Islands: the subject position in English is by default specific, so it is possible that Subject Islands are unique because they are specific.

Participants In order to keep the number of items per survey manageable, the Island types were split among two experiments: experiment 1 tested Adjunct, CNPC, Subject, and Whether Islands, while experiment 2 tested Specificity and Relative Clause Islands. 21 University of Maryland undergraduates participated in experiment 1, and 22 Princeton University undergraduates participated in experiment 2.

Design

8 lexicalizations of each condition were created and distributed among 8 lists using a Latin Square design. Conditions from an unrelated experiment were added to both experiments for a total of 26 items in experiment 1 and 28 items in experiment 2. Three orders for each list were created by pseudorandomizing the conditions such that no two related conditions were consecutive, for a total of 24 surveys for each experiment. The task for both experiments was magnitude estimation, and the reference sentence was *What did you ask if your mother bought for your father?*. The instructions were a modified version of the instructions published with the WebExp software suite (Keller et al. 1998).

Results

Responses were divided by the score of the reference sentence and log-transformed prior to analysis. The means and standard deviations for each level of each factor for all of the Island types is summarized in the following table:

Table 6.2. Wh-in-situ: descriptive results

wh-in-situ:	no wh-in-situ				wh-in-situ			
structure:	non-island		island		non-island		island	
	mean	SD	mean	SD	mean	SD	mean	SD
Adjunct	.24	.21	.22	.25	.02	.17	-.04	.27
Subject	-.03	.25	-.06	.29	-.04	.21	-.21	.40
CNPC	.28	.23	.27	.23	.07	.23	-.03	.26
Whether	.32	.25	.18	.16	.08	.25	-.02	.36
Specificity	.32	.25	.35	.22	-.04	.28	-.11	.43
Rel. Clause	.33	.19	.38	.23	0.0	.21	.05	.17

Two-way repeated measures ANOVAs were performed on each Island type. The results are summarized in the following table:

Table 6.3. Wh-in-situ: Two-way repeated measures ANOVA

	STRUCTURE			WH-IN-SITU			STRUC x WH		
	F	p	eta^2	F	p	eta^2	F	p	eta^2
Adjunct	1.3	.267	—	39.1	$***$	.662	0.4	.534	—
Subject	9.1	$**$	.312	2.1	.160	—	5.8	$*$	.221
CNPC	5.4	$*$	.212	23.1	$***$	.536	2.2	.153	—
Whether	11	$***$	.354	14.3	$***$	.417	0.3	.604	—
Specificity	0.2	.674	—	45.6	$***$	.685	1.1	.298	—
Relative Clause	3.8	.065	—	71.6	$***$	.773	0.0	.969	—

$*** = p < .001, ** = p < .01, * = p < .05$

While there were various significant main effects of STRUCTURE and WH-IN-SITU, the focus of this experiment was on the interaction of the two, as this indicates an Island effect with wh-in-situ. As the ANOVA table indicates, the only Island type to show a significant interaction is Subject Islands, with a small to medium sized effect.

Discussion

The experiments presented in this subsection were designed to overcome the shortcomings of the first experiment, and determine whether there is indeed a wh-in-

situ Island effect for several Island types. The results suggest that there is no wh-in-situ Island effect in English for Adjunct, CNPC, Whether, Specificity, and Relative Clause Islands. However, there is a wh-in-situ Island effect for Subject Islands. This result clarifies the lack of significant difference between overt movement and wh-in-situ with Subject Islands in the first experiment: there is an Island effect for both types of dependency. Unfortunately, because of the relative nature of magnitude estimation judgments it is not clear whether wh-in-situ in an Island structure, especially the Subject Island, is considered categorically acceptable or unacceptable. The follow-up experiment in the following subsection investigates that very question.

6.1.3 Categorical acceptability of overt and covert movement Islands

The design for this experiment is very similar to the first experiment testing Huang's (1982) comparison of overt movement out of an Island to wh-in-situ in an Island. 24 University of Maryland undergraduates were presented with 2 tokens each of overt movement and covert movement out of the following Islands: Adjunct, CSC, CNPC, Relative Clause, Subject and Whether Islands. 12 tokens of each condition were created and distributed among 6 lists (2 per list). 4 orders of each list were created for 24 lists. 20 acceptable fillers were added to the lists to better approximate a 1:1 ratio of acceptable to unacceptable items, for a total of 44 items. The task was a categorical yes/no task.

Results and Discussion

Since each participant judged 2 tokens of each condition, there were three possible response patterns, two unambiguous and one ambiguous: both tokens judged *yes*, both *no*, or one of each judgment. The total number of each type of unambiguous judgment was summed across participants and compared using a Sign Test to determine the categorical acceptability of each condition. Ambiguous responses were excluded from the analysis:

Table 6.4. Categorical acceptability of overt movement and wh-in-situ in Islands

	overt movement				wh-in-situ			
	yes	no	p	category	yes	no	p	category
Adjunct	1	19	.001	no	6	13	.167	—
CSC	1	17	.001	no	4	15	.019	no
CNPC	1	19	.001	no	5	12	.142	—
Rel Cl.	1	22	.001	no	6	12	.238	—
Subject	1	20	.001	no	3	15	.008	no
Whether	6	15	.078	—	7	14	.189	—

As the table indicates, all of the overt movement Island types were significantly judged as categorically unacceptable except for Whether Islands, which were marginally significant. Wh-in-situ in Islands, on the other hand, were less straightforward. The only clear cases were CSC and Subject Islands, which were judged as categorically unacceptable. None of the other Island types reached significance, although at least two-thirds of participants judged each Island as unacceptable, suggesting that increasing the sample size would lead to a significant number of unacceptable responses.

Focusing momentarily on Subject Islands, this follow-up suggests that the relative wh-in-situ Subject Islands detected by magnitude estimation do indeed lead

to a categorical judgment of unacceptable. This in turn strongly suggests that wh-in-situ Subject Islands are ungrammatical. However, this result is a double-edged sword: if the other wh-in-situ Island types are also categorically unacceptable, as suggested by the preponderance of *no* judgments, then we will be left with a minor mystery as to the source of the unacceptability as there were no detectable wh-in-situ Island effects in the previous studies other than the Subject Island. It may be the case that bi-clausal (all of the Island structures involve two clauses) multiple wh-questions are independently unacceptable for a majority of the participants. This suggests that an in-depth follow-up study on the categorical acceptability of multiple-wh questions may be in order in the future.

6.1.4 Theoretical implications of Subject Island effects with wh-in-situ

The results of the three studies presented in this section can be summarized by the following set of claims:

1. Wh-in-situ within an Island is significantly more acceptable than overt wh-movement out of an Island for many Island types.

2. Wh-in-situ within a Subject Island is not significantly different than overt wh-movement out of a Subject Island.

3. Wh-in-situ within an Island is not significantly different than wh-in-situ within a minimally different non-Island.

4. Wh-in-situ within a Subject Island is significantly less acceptable than wh-in-situ within a minimally different non-Island.

5. Wh-in-situ within a Subject Island is judged as categorically unacceptable. The other Island types are judged unacceptable by a majority of participants, but not enough to reach significance.

Or, in other words, there are Subject Island effects with wh-in-situ in English that are nearly identical to overt wh-movement Subject Islands. These results have consequences for syntactic theory on at least two levels. First, at the level of individual analyses, these results suggest that Subject Islands are unique among the other Islands tested, at least with respect to wh-in-situ. This raises obvious problems for analyses in which the underlying cause of Subject Island effects is the same as other Island types, such as the Subjacency approach to Islands (Chomsky 1973, 1986), or the CED (Condition on Extraction Domains) approach of Huang 1982 (see Stepanov 2007 for other arguments against the CED approach). At the level of analysis types, such as covert movement analyses versus choice-function analyses, these results change the empirical landscape. These analyses must be modified to allow the possibility of wh-in-situ Island effects, but restrict this to Subject Island effects. This entails a major modification to in-situ based approaches such as the choice-function approach in Reinhart 1997 in which Island effects are restricted to overt movement.

Perhaps more importantly, these results suggest a new parallelism between overt wh-movement and wh-in-situ in that both exhibit at least one Island effect. This could be interpreted as a new argument for a movement-based approach to wh-

in-situ, either through covert movement or through overt movement with Spell-out of the lower copy. In fact, a movement-based approach also offers the possibility of accounting for the unique nature of Subject Islands through freezing-style approaches to Subject Islands (Wexler and Culicover 1981): if subjects must move from a VP internal position to the specifier of IP in English, and if movement of a phrase causes that phrase to become an Island to further movement, then the Island status of subjects for wh-in-situ would follow directly from movement. As Cedric Boeckx points out (p.c.), this predicts that wh-in-situ Island should not occur if the subject has not moved, as is possible in some Romance languages (e.g., Spanish in Gallego 2007. The appropriate follow-up study is left to future research.

6.2 Distance effects and wh-in-situ in English

The previous section used various experimental techniques to uncover a new parallelism in acceptability between overt wh-movement and wh-in-situ in English: the presence of Subject Island effects. Those results suggest that wh-in-situ may be more like overt wh-movement than previously thought. This section reports a second study investigating another property of overt wh-movement - a direct relationship between the distance of movement (in clauses) and acceptability - and whether wh-in-situ shows similar effects.

6.2.1 Distance and wh-dependencies

In a study of the processing of wh-dependencies and the effect of dependency length on working memory, Phillips et al. 2005 report the results of an offline rating study in which participants were asked to rate the *complexity* of sentences along a 5-point scale. They found that manipulating the length of an overt wh-movement dependency affected complexity ratings such that longer wh-dependencies were rated more complex:

(47) The detective hoped [that the lieutenant knew [**which accomplice** the shrewd witness would recognize __ in the lineup]].

(48) The lieutenant knew [**which accomplice** the detective hoped [that the shrewd witness would recognize __ in the lineup]].

Table 6.5. Mean complexity ratings for short and long movement, from Phillips et al. 2005

	mean	SD
Short – 1 clause	2.71	0.65
Long – 2 clauses	3.51	0.51

The first question is whether this effect arises in acceptability tasks as well, and if so, whether wh-in-situ, with no visible movement, is also affected by the distance of the dependency.[1] This can be straightforwardly tested with materials similar to those in Phillips et al. 2005:

[1] An interesting question is whether participants can actually differentiate between complexity and acceptability. Given the likelihood that complexity influences acceptability judgments and vice versa, without a comprehensive investigation of the sensitivity of participants to both properties it is difficult to interpret the course of the either a complexity effect or an acceptability effect.

(49) Overt wh-movement distance

a. Who hoped that you knew **who** the mayor would honor __?

b. Who knew **who** hoped that the mayor would honor __?

The second set crucially manipulated the distance of the covert movement dependency:

(50) Wh-in-situ distance

a. Who hoped that you knew **who** would honor **who**?

b. Who knew **who** hoped that you would honor **who**?

However, because the manipulation of distance in (50) is necessarily confounded with the size of the embedded question (1-clause covert movement involves a 1-clause embedded question, 2-clause covert movement has 2-clause embedded question), a third minimal pair is necessary to tease apart the contribution of covert movement distance and embedded question size. In these two conditions, the distance of the wh-in-situ dependency is always the entire length of the question (answers to these questions are pair lists involving both the matrix wh-word and the in-situ wh-word, each of which is marked in **bold**), but the size of the embedded question is manipulated (the embedded question begins with *whether* and is marked in *italics*):

(51) Embedded question size control

a. **Who** hoped that you knew *whether the mayor would honor* ***who***?

b. **Who** knew *whether you hoped that the mayor would honor* ***who***?

If any effect found for (50) is due to the size of the embedded question, we would expect the same effect for these conditions because the size of the embedded question is manipulated in these conditions; if the effect is due to the distance of the wh-in-situ dependency then we would not expect an effect for these conditions because the distance of the wh-in-situ dependency is held constant.

Participants

26 University of Maryland undergraduates participated in this experiment. All were self-reported monolingual, native speakers of English. Participants were paid for their participation.

Design

8 lexicalizations of each of the 6 conditions were created and combined with 8 lexicalizations of 26 conditions from an unrelated experiment, then distributed among 8 lists using a Latin Square design. 4 pseudorandomized orders of each list were created such that related conditions were never consecutive. The task was magnitude estimation. The instructions were based upon the published instructions in the WebExp software suite (Keller et al. 1998), and the reference sentence was *What did you ask if your mother bought for your father?*

Results and Discussion

Responses were divided by the score of the reference sentence and log-transformed prior to analysis. Paired t-tests were performed on each condition set (all p-values are two tailed):

Figure 6.2. Effects for wh-movement distance

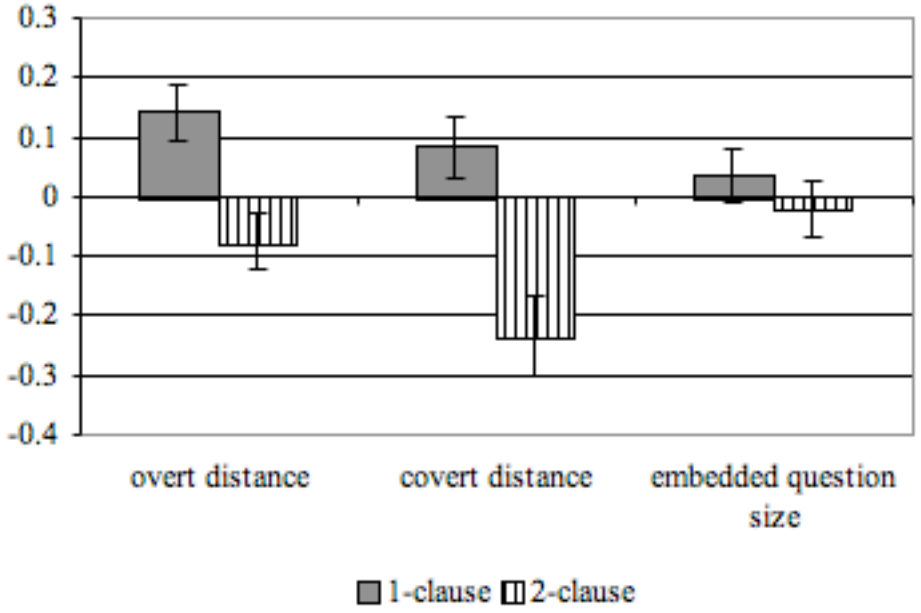


Table 6.6. Paired t-tests for wh-movement distance conditions

	Short		Long					
	Mean	SD	Mean	SD	*df*	t	*p*	effect size (r)
overt distance	0.14	0.24	-0.07	0.24	25	3.99	.001	.62
in-situ distance	0.08	0.26	-0.23	0.34	25	3.83	.001	.61
embedded size	0.04	0.23	-0.02	0.24	25	0.88	.388	—

There were large significant effects for distance with overt movement and wh-in-situ, with shorter dependencies being more acceptable than longer dependencies, but no effect of embedded question size. This suggests that the distance effects with wh-in-situ are indeed due to the dependency length and not due to the size of the embedded question. These results suggest yet another parallelism between overt wh-movement and wh-in-situ dependencies: the effect of distance. However, as the next subsection will discuss, there are potential analyses of the wh-in-situ distance effect that do not

involve movement.

Before beginning that discussion, it is worth mentioning that if the wh-in-situ length effect does turn out to be due to (covert) movement, the lack of effect of embedded question size would become evidence that a large scale pied-piping analysis along the lines of Nishigauchi 1990 and Dayal 1996 is incorrect for English Whether Islands. Under a large scale pied-piping analysis, the lack of Whether Island effects with wh-in-situ follows from covert movement of the entire Whether Island to the matrix C: because the in-situ wh-word does not move out of the Island, there is no Island effect. An interesting side effect of this analysis is that the larger the constituent being moved, the shorter the movement distance. In this case, large scale pied-piping of the entire embedded question would mean that the larger embedded question would only covertly move one clause, whereas the smaller embedded question must move two clauses. If covert movement leads to distance effects, the pied-piping analysis would predict that larger embedded questions should be more acceptable than smaller embedded questions because smaller embedded questions must move farther.[2] However, there was no effect of the embedded question size in this experiment.

[2]It is also possible that movement of larger constituents leads to a decrease in acceptability that neutralizes the benefit larger constituents gain from moving shorter distances. Such a counteranalysis would require a detailed investigation of the effect of constituent size on the effect of acceptability that is beyond the scope of this dissertation.

6.2.2 Distance and Binding dependencies

If distance effects are a characteristic property of movement, the distance effect with wh-in-situ is compelling evidence for a movement approach. One possibility is that the successively cyclic nature of movement, which entails one instance of the movement operation for each CP crossed, has a direct effect on acceptability. Such an analysis would mean that either covert movement is successive cyclic contrary to accepted wisdom (see especially Epstein 1992), or that all movement is overt movement, with the option of pronouncing the lower copy (e.g., Bošković 2002).

There are, of course, other plausible explanations for the distance effect that have nothing at all to do with the syntactic operations involved. For instance, the Phillips et al. (2005) study suggests that the farther the displaced wh-word is from the gap position, the harder it is to process, perhaps because the representation of the wh-word in working memory decays over time. This working memory cost for long distance dependencies could underlie the distance effect for overt wh-movement. For wh-in-situ dependencies there is no displacement, so the effect of working memory is less apparent. However, we know that the interpretation of wh-in-situ is dependent upon the matrix wh-word, therefore a representation of the first wh-word must be maintained in memory in order to interpret the second wh-word. This is very similar to the memory requirement of overt displacement, hence it may not be surprising to find a wh-in-situ distance effect.

An analysis such as the one above in which the interpretation of two functionally related items leads to a processing cost on acceptability that is distance dependent

makes a very specific prediction: distance effects should arise with other dependencies that lead to this functional interpretation, not just wh-dependencies. As such, three binding dependencies were tested for distance effects: a bound-variable relationship between a wh-word and a possessive pronoun, a bound variable relationship between a quantifier and a possessive pronoun, and a general binding relationship between an R-expression and a possessive pronoun.[3]

(52) Bound variable, wh-word

i. Who hoped that you knew who found their wallet?

ii. Who knew who hoped that the police found their wallet?

(53) Bound variable, quantifier

i. Who hoped that you knew if everyone found their wallet?

ii. Who knew if everyone hoped that the police found their wallet?

(54) Coreference, R-expression

i. Who hoped that you knew if John found his wallet?

ii. Who knew if John hoped that the police found his wallet?

Participants

22 Princeton University undergraduates participated in this experiment. All were self-reported native speakers of English. All volunteered their time for this study.

[3]Possessive pronouns were chosen to avoid Principle B violations in the short distance conditions with standard pronouns, or a Principle A violation in the long distance conditions with reflexives.

Design

8 lexicalizations of each condition were constructed and combined with 8 lexicalizations of 16 other conditions from an unrelated experiment, as well as 6 filler items to balance acceptability, then distributed among 8 lists using a Latin Square design. 3 orders of each list were created by pseudorandomizing such that no two related conditions were consecutive, for a total of 24 lists.

The task was magnitude estimation. The instructions were a modified version of the instructions distributed with the WebExp software suite (Keller et al. 1998). The reference sentence was: *What do you ask if your mother bought for your father?*

Results and Discussion

Responses were divided by the score of the reference sentence and log-transformed prior to analysis. Paired t-tests were performed on each pair of conditions; all p-values are two-tailed:

Figure 6.3. Results for binding distance

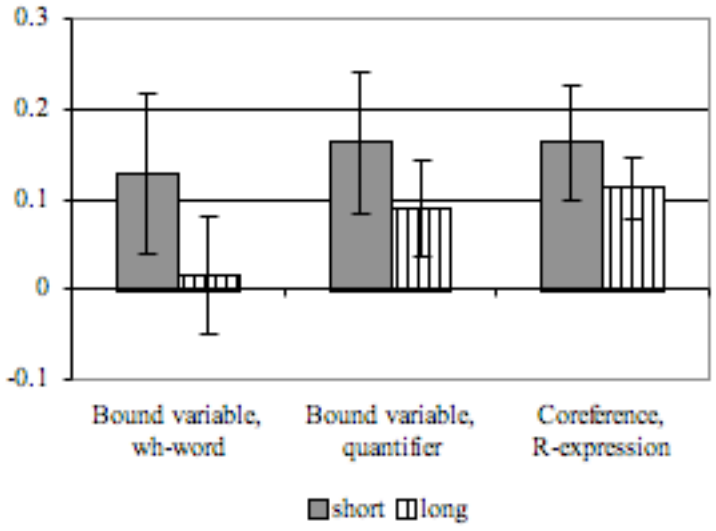


Table 6.7. Results for binding distance t-tests

	Short		Long				
	Mean	SD	Mean	SD	*df*	t	*p*
BV, wh-word	0.13	0.41	0.02	0.30	21	1.386	.180
BV, quantifier	0.16	0.36	0.09	0.25	21	0.899	.379
R-expression	0.16	0.30	0.11	0.16	21	0.899	.379

There were no significant effects of distance for any of the binding dependencies tested. There is a non-significant trend in the correct direction (lower acceptability with longer distance) for all three dependency types, which raises the question of whether a larger sample size would lead to significant effects. A power analysis suggests that if the means and standard deviations are accurate, sample sizes of 53, 106, and 144 respectively would be needed for the results to become significant. Given that the wh-distance results appeared with a sample size of 26, the relatively weak effect of distance for binding dependencies suggests that even if significant effects could be obtained, they would likely have a different underlying cause.

The lack of distance effects with binding dependencies suggests that the distance effects with wh-in-situ dependencies is not due to the interpretive relationship between the two wh-words. While acceptability experiments cannot rule out a processing explanation, these results do indicate that a processing explanation must take into account the differential effects between wh-in-situ dependencies and binding dependencies; a broad stroke analysis based on working memory cannot capture this distinction.

While these results once again suggest a strong similarity between overt movement and wh-in-situ dependencies, they also have interesting consequences for recent

attempts to unify binding and wh-dependencies (e.g., Hornstein 2000). While it is still possible that binding and wh-dependencies are built by the same structure building operation (perhaps movement), such unificational theories face the same problem as processing-based accounts of the distance effects: they must distinguish between the distance sensitivity of wh-dependencies and the insensitivity of binding dependencies. There are many possibilities (perhaps binding dependencies are not successive cyclic, perhaps A-movement is insensitive to distance effects, etc.) awaiting future research.

6.3 Refining the analyses of wh-in-situ in English

This chapter began with the general consensus that acceptability judgments had little contribute to the debate between competing analyses of wh-in-situ in English. The studies presented in this chapter have uncovered new acceptability effects that bear directly on the debate:

1. There are Subject Island effects with wh-in-situ in English.

2. There are distance effects with wh-in-situ in English.

3. There are no effects of the size of Whether Islands.

4. There are no distance effects with binding in English.

Or, in other words, wh-in-situ shows two characteristic properties of movement that were previously unnoticed: distance effects and at least one island effect. These results strongly suggest a movement style approach to wh-in-situ in English. Furthermore, they argue against one of the movement approaches: the large scale pied-piping ap-

proach of Nishigauchi 1990 and Dayal 1996. These results also raise many interesting questions for future research, although the exact nature of the questions depends on the movement approach.

Under a dual cycle syntax model in which covert movement is possible, the results of these experiments could be interpreted as evidence for covert movement. In particular, the Subject Island effects with covert movement could receive a straightforward account under a freezing-style analysis (Wexler and Culicover 1981), although not under a linearization and freezing analysis such as the one in Uriagereka 1999 (though the other Island effects could still be analyzed as deriving from linearization as they do not constrain covert movement). Such an analysis also raises the possibility that the distance effect with covert movement is due to successive cyclicity, which runs counter to the general consensus in the field that covert movement is not successive cyclic (see especially Epstein 1992). Furthermore, this would raise the question of why overt movement is required to be successively cyclic while covert movement can move in one fell swoop to escape Wh-Islands.

As Norbert Hornstein (p.c) points out, a single cycle syntax model avoids the problem of positing successive cyclic covert movement by eliminating covert movement altogether. Under such a model, the distance effects with overt wh-movement and wh-in-situ derive from the same source: overt movement. The difference between the two dependencies rests solely in which copy, higher or lower, is pronounced (e.g., Bošković 2002). Such a model can also adopt the freezing-style analysis for Subject Islands, but must also eschew the linearization analysis for Subject Islands.

These facts are also compatible with an unselective binding approach with null operator movement, such as the one in Tsai 1994. Under this model, wh-in-situ does involve movement, but it is (overt) movement of a null operator, not of the wh-word. The distance facts fall out naturally from the movement of the null operator. The Subject Island effect can also be attributed to a freezing-style analysis, especially if the null operator must move out of the Island via something like short movement from Hagstrom 1998: because subjects have already moved, the null operator cannot escape the Island.

The facts in this chapter raise the most difficult problems for AGREE based (Chomsky 2000, 2001, 2005) and choice-function based analyses (Reinhart 1997). Both types of analyses define Island effects as reflexes of movement, therefore they must be weakened to account for the Subject Island effects with wh-in-situ. Furthermore, while both analyses involve a long distance dependency (long distance AGREE and existential closure respectively), neither is specific to wh-dependencies. The lack of distance effects with binding dependencies suggests that distance effects may be unique to wh-dependencies. If AGREE or existential closure were causing the distance effects, we would not expect them to arise only in wh-dependencies.

Chapter 7

Conclusion

This dissertation has argued that the tools of experimental syntax can be used to further our understanding of the relationship between the nature of acceptability and the nature of grammatical knowledge. To that end, several existing claims about the nature of grammaticality were investigated:

1. That grammatical knowledge is gradient.

2. That grammatical knowledge encodes a distinction between constraints that are affected by context and those that are not.

3. That the stability or instability of violations reflects a difference in the nature of the grammatical knowledge underlying the violation.

4. That acceptability judgments are affected by processing effects.

5. That acceptability judgments have nothing further to contribute to determining the number or nature of wh-in-situ dependency forming operations.

The results of those investigations suggest that:

1. While grammatical knowledge may still be gradient, the categorical distinction between grammatical and ungrammatical is also real, as it arises (unprompted) even in continuous tasks like magnitude estimation .

2. While context may have an effect on some acceptability judgments, it is likely that those effects reflect non-structural constraints such as pragmatics or information structure. Furthermore, there is no evidence that context affects the wh-movement properties investigated in this dissertation: Island effects.

3. While there still may be different causes underlying various violations, satiation is not likely to differentiate among them, as satiation is most likely an artifact of experimental design, and not a property of acceptability judgments.

4. While it goes without saying that processing effects affect acceptability judgments, it is not the case that *all* processing effects have an effect. This differential sensitivity suggests a methodology for validating the possibility of processing-based explanations of acceptability effects: first investigate whether the processing effects in question have an effect on acceptability.

5. While the value of non-acceptability data such as possible answers are undoubtedly valuable for investigations of wh-in-situ, two new acceptability-based effects were presented that may have important consequences for wh-in-situ theories. While some hypotheses were suggested, future research is necessary to identify the relationship between those effects and grammatical knowledge.

Like many studies, the work presented in this dissertation raises far more questions than it answers. However, it is clear from these results that there is a good deal of potential for experimental syntax to provide more than a simple fact-checking service for theoretical syntax: the tools of experimental syntax are in a unique position to further our understanding of the relationship between acceptability and grammatical

knowledge, and ultimately, refine our theories of the nature of grammatical knowledge itself.

Bibliography

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72:32–68.

Bošković, Željko. 2002. On multiple wh-fronting. *Linguistic Inquiry* 33:351–383.

Bresnan, Joan. 2007. A few lessons from typology. *Linguistic Typology* 11.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, Massachusetts: M.I.T. Press.

Chomsky, Noam. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, ed. Stephen Anderson and Paul Kiparsky, 232–286. Holt, Rinehart and Winston.

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

Chomsky, Noam. 2000. Minimalist inquiries: The framework. In *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*, ed. Roger Martin, David Michaels, and Juan Uriagereka, 89–156. MIT Press.

Chomsky, Noam. 2001. Derivation by phase. In *Ken Hale: A life in linguistics*, ed. Michael Kenstowicz, 1–52. Cambridge, Massachusetts: MIT Press.

Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36:1–22.

Cohen, Jacob. 1973. Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement* 33:107–112.

Conover, W. J., and R. L. Iman. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician* 35:124–129.

Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Sage Publications.

Crain, Stephen, and Janet Fodor. 1987. Sentence matching and overgeneration. *Cognition* 26:123–169.

Crain, Stephen, and Rosalind Thornton. 1998. *Investigations in Universal Grammar: A guide to experiments on the acquisition of syntax and semantics*. Cambridge, Massachusetts: The MIT Press.

Dayal, Vaneeta. 1996. *Locality in [w]h-quantification: [q]uestions and relative clauses in [h]indi*. Kluwer.

Dayal, Vaneeta. 2006. Multipl [w]h-questions. In *The syntax companion*, ed. M. Everaert and H. van Riemsdijk, chapter 44. Blackwell.

Deane, Paul. 1991. Limits to attention: A cognitive theory of island phenomena. *Cognitive Linguistics* 2:1–63.

Edelman, Shimon, and Morten Christiansen. 2003. How seriously should we take minimalist syntax? *Trends in Cognitive Science* 7:59–60.

Epstein, Samuel David. 1992. Derivational constraints on A'-chain formation. *Linguistic Inquiry* 23:235–260.

Erteschik-Shir, Nomi. 2006. What's what? In *Gradience in grammar*, ed. Caroline Fery, Fanselow Gisbert, Matthias Schlesewsky, and Ralf Vogel. Oxford University Press.

Fanselow, G., and S. Frisch. 2004. Effects of processing difficulty on judgments of acceptability. In *Gradience in grammar*, ed. Fery C. Schlesewsky M. Vogel R. Fanselow, G. Oxford University Press.

Featherston, Sam. 2005a. Magnitude estimation and what it can do for your syntax: [s]ome wh-constraints in [g]erman. *Lingua* 115:1525–1550.

Featherston, Sam. 2005b. Universals and grammaticality: wh-constraints in German and English. *Linguistics* 43:667–711.

Fisher, Ronald. 1922. On the interpretation of Chi-squared from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85:87–94.

Frazier, L., and G. Flores d'Arcais. 1989. Filler driven parsing: A study of gap filling in Dutch. *Journal of Memory and Language* 28:331–344.

Frazier, Lyn, and Jr. Clifton, Charles. 2002. Processing "d-linked" phrases. *Journal of Psycholinguistic Research* 31:633–659.

Gallego, Angel. 2007. Phase theory and parametric variation. Doctoral Dissertation, Univeritat Autonoma de Barcelona.

Goodall, Grant. 2005. Satiation and inversion in wh-questions. Talk given at University of Hawaii.

Hagstrom, Paul. 1998. Decomposing questions. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Hiramatsu, Kazuko. 2000. Accessing linguistic competence: Evidence from children's and adults' acceptability judgments. Doctoral Dissertation, University of Connecticut.

Hornstein, Norbert. 2000. *Move! a minimalist theory of construal*. Blackwell.

Huang, C.-T. 1982. Move wh in a language without wh-movement. *The Linguistic Review* 1:369–416.

Kaan, Edith, and Laurie Stowe. ms. *Developing an experiment*.

Keller, Frank. 2000. Gradience in grammar: Experimental and computational aspects of degress of grammaticality. Doctoral Dissertation, University of Edinburgh.

Keller, Frank. 2003. A psychophysical law for linguistic judgments. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*.

Keller, Frank, M. Corley, S. Corley, L. Konieczny, and A. Todirascu. 1998. Webexp. Technical report hcrc/tr-99, Human Communication Research Centre, University of Edinburgh.

Kuno, Susumu. 1976. Gapping: a functional analysis. *Linguistic Inquiry* 7:300–318.

Levene, H. 1960. Robust tests for equality of variances. In *Contributions to probability and statistics*, ed. I. Olkin, 278–292. Stanford University Press.

Lodge, Milton. 1981. *Magnitude scaling: Quantitative measurement of opinions*. Sage.

Lorch, R. F., Jr., and J. L. Myers. 1990. Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognitition* 16:149–157.

Myers, James. 2006. Minijudge: Software for minimalist experimental syntax. In *Proceedings of ROCLING 18*, 271–285.

Nishigauchi, Taisuke. 1990. *Quantification in the theory of grammar*, volume 37. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pesetsky, David. 1987. Wh-in-situ: Movement and unselective binding. In *The representation of (in)definiteness*, ed. Eric J. Reuland and Alice G. B. ter Meulen, 98–129. Cambridge, Massachusetts: MIT Press.

Phillips, Colin, Nina Kazanina, and Shani Abada. 2005. ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research* 22:407–428.

Pickering, Martin, and Michael Traxler. 2003. Evidence against the use of subcategorization frequency in the processing of unbounded dependencies. *Language and Cognitive Processes* 18:469–503.

Reinhart, Tanya. 1997. Quantifier scope: How labor is divided between QR and choice functions. *Linguistics and Philosophy* 20:335–397.

Ross, John. 1967. Constraints on variables in syntax. Doctoral Dissertation, Massachusetts Institute of Technology.

Sag, Ivan, Inbal Arnon, Bruno Estigarribia, Philip Hofmeister, T. Florian Jaeger, Jeanette Pettibone, and Neal Snider. submitted. Processing accounts for superiority effects .

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

Seaman, J. W., S. C. Walls, S. E. Wide, and R. G. Jaeger. 1994. Caveat emptor: Rank transform methods and interactions. *Trends in Ecological Evolution* 9:261–263.

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

Sorace, Antonella, and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115:1497–1524.

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

Stevens, Stanley Smith. 1957. On the psychophysical law. *Psychological Review* 64:153–181.

Tsai, Wei-Tien Dylan. 1994. On nominal islands and LF extraction in Chinese. *Natural Language and Linguistic Theory* 12:121–175.

Wexler, Kenneth, and Peter Culicover. 1981. *Formal principles of language acquisition*. Cambridge, Massachusetts: MIT Press.

Wilcox, Rand R. 1997. *Introduction to robust estimation and hypothesis testing*. Academic Press.