

Chapter 2

Revisiting Satiation

This chapter investigates a phenomenon, known in the literature as *syntactic satiation*, in which judgments of certain violations appear to get better, that is more acceptable, after several repetitions. Satiation serves as a good starting point for building our research program because it has consequences at each of the three levels laid out in chapter 1: At a theoretical level, instability in acceptability judgments threatens to undermine the syntactic analyses that are built on acceptability data; at a methodological level, satiation has been proposed as test for whether violations are grammar-based or processing-based; and at an empirical level, determining the factors that affect the stability of acceptability is a crucial step in laying the foundation for a research program that relies on acceptability judgments. What we shall see, however, is that the apparent instability of acceptability is actually an artifact of informal judgment collection techniques. This is welcome news for syntactic analyses built on those judgments, and for research programs such as the one laid out in chapter 1 that relies on acceptability judgments. The downside of this finding, however, is that satiation can not serve as a simple test for determining whether violations are grammar-based or processing-based.

2.1 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) - a possibility that is intriguing from the point of view of comparing grammar-based and processing-based analyses of acceptability facts. Whatever the interpretation of satiation, Snyder argues that if it systematically occurs with some violations and not others, then it isn't 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 En-

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 2.1. 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 lead 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

¹Hiramatsu 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

Table 2.2. Summary of results for Snyder 2000, Hiramatsu 2000, and Goodall 2005

Sentence Type	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, however, 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 whether additional exposures would lead to satiation of Subject Islands (which were marginal in 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).

in an attempt to isolate the conditions necessary to license the type of judgment instability 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.

2.2 The replication problem

2.2.1 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: subjects 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.

2.2.1.1 Subjects

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.

2.2.1.2 Materials and Design

The materials and design for the direct replication were those from Snyder 2000. The materials for the replication with confidence judgments were also those from Snyder 2000, but there was the additional confidence task with each judgment, as well

²The original materials were graciously provided by William Snyder.

as an additional paragraph in the instructions explaining the confidence judgment.

The modified replication followed the general design of Snyder 2000, but included a few minor modifications. First, there were 8 violations per block instead of 7, which means there were only 2 acceptable sentences per block instead of 3 (the logic of this modification will be apparent in section 2.3). Second, the violations were all Island violations, whereas the Snyder 2000 violations included non-Island violations such as the That-trace effect and the Want-for effect. 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 (except for LBC violations for which this impossible); iii) all 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 orders, rather than the 2 orders from a forward-backward counterbalance.

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 2.3. 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?

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

Table 2.4. Results from the direct replication

Violation	No to Yes	Yes to No	<i>p</i> – <i>value</i>
Adjunct Island	5	6	N/A
CNPC Island	2	2	N/A
LBC Violation	0	0	N/A
Subject Island	5	2	.45
That-trace effect	5	3	.73
Want-for effect	2	2	N/A
Whether Island	3	4	N/A

Table 2.5. Results from the replication with confidence

Violation	No to Yes	Yes to No	<i>p</i> – <i>value</i>
Adjunct Island	4	4	N/A
CNPC Island	2	0	N/A
LBC Violation	1	1	N/A
Subject Island	5	1	.22
That-trace effect	7	2	.18
Want-for effect	5	3	.73
Whether Island	6	5	N/A

Table 2.6. Results from the modified replication

Violation	No to Yes	Yes to No	<i>p</i> – <i>value</i>
Adjunct Island	4	1	.38
CNPC Island	3	1	.63
Coordinate Structure Constraint	3	4	N/A
Infinitival Sentential Subject Island	1	1	N/A
LBC Violation	2	3	N/A
Relative Clause Island	3	2	N/A
Sentential Subject Island	3	0	.25
Whether Island	4	4	N/A

2.2.1.4 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. Furthermore, the replication problem cannot be attributed to population differences as students from both private and public universities are represented in the replications:

Table 2.7. Summary of Snyder 2000 and 5 subsequent replication attempts

Study:	Snyder	Hiramatsu	Goodall	Direct	Confidence	Modified
School:	MIT	UConn	UCSD	Penn	UMD	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

2.2.2 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 licenses of 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:

1. the design
2. the task
3. the definition of the effect

The goal of experimental design is to control for any task or subject 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 subjects (if Subject 1 sees sentence A first, then Subject 2 sees sentence A last). However, these experiments do not control for the possibility that subjects 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 subjects toward judging sentences as acceptable later in the experiment in an attempt to balance their responses.³

The task chosen for these experiments was the *yes/no task* in which subjects 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 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 subjects confirms this as nearly every subject 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 subjects 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 subject
2. Count the number of *yes* responses after the last two exposures of each violation for each subject
3. If the number of *yes* responses in the last two exposures is higher than in the first two, the subject is defined as satiating for that violation
4. If the number of *yes* responses in the last two exposures is lower than in the

³One anonymous LI reviewer suggests that the unbalanced nature of the composition ensures that subjects 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 subject that may be memorizing their responses.

first two, the subject is defined as not-satiating for that violation

5. For each violation, count the number of subjects that satiated and the number of subjects that not-satiated (n.b. the subjects whose responses were stable are ignored)
6. If there are statistically more satiators than non-satiators then the violation is said to satiate.

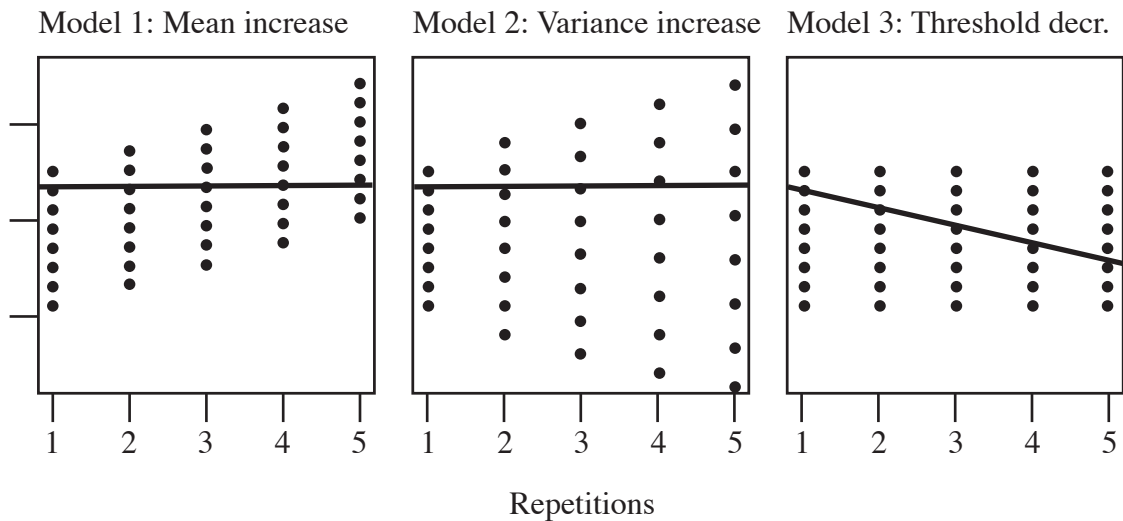
Basically, this definition asks: For those people who have unstable judgment, are more of them unstable in a positive direction or in a negative direction. While this a possible topic of investigation, because of the elimination of stable subject from the analysis, this definition artificially limits 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 subjects 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 subjects are excluded, the final analysis is conducted over those subjects who demonstrated instability, or in other words, subjects 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 com-

paring the number of subjects 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 do subjects sometimes adopt this response strategy for Whether Islands but never adopt it for 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 isn't 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 actually 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 2.1. Three models of satiation



The first model is one in which the acceptability judgments are increasing over time, and eventually those that started below the yes/no threshold 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 doesn't 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.⁴

2.2.3 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:

⁴In 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 decrease over the course of these experiments, which could be caused by a change in the category threshold.

Table 2.8. 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, since 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 a statistical definition that can be argued over, and *instability*, which is any statistically definable change in judgments within a single violation type. Since any instability is methodologically interesting, i will use the terms *stability/instability* or *stable/unstable* to refer to this change in contrast to *satiation* where appropriate. Because of the potential disagreement about the definition of satiation, this chapter more accurately focuses on stability rather than satiation, in that the goal is to identify factors that cause any changes in judgment (be it violation type x time, or task x design), and to quantify those changes to determine if they bear on linguistic theories.

2.3 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

Table 2.9. 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
<i>CNPC Island</i>	<i>CNPC Island</i>
<i>Whether Island</i>	<i>Whether Island</i>
Acceptable Sentence	Adjunct Island
Acceptable Sentence	Relative Clause 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:

2.3.1 Subjects

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. 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.

2.3.2 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 ac-

Table 2.10. Results from the unbalanced yes/no task

Violation	No to Yes	Yes to No	<i>p</i> – <i>value</i>
Adjunct Island	4	1	.38
Coordinate Structure Constraint	3	4	N/A
Infinitival Sentential Subject Island	1	1	N/A
LBC Violation	2	3	N/A
Relative Clause Island	3	2	N/A
Sentential Subject Island	3	0	.25
CNPC Island	3	1	.63
Whether Island	4	4	N/A

Table 2.11. Results from the balanced yes/no task

Violation	No to Yes	Yes to No	<i>p</i> – <i>value</i>
Adjunct Island	0	0	N/A
Relative Clause Island	2	0	N/A
CNPC Island	0	0	N/A
Whether Island	0	2	N/A

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

2.3.3 Results

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

Applying a basic definition of instability, for instance counting the number of

subjects whose judgments change even once, we find that 15 subjects' judgments change in the unbalanced experiment, and 3 subjects' judgments change in the balanced experiment. A Chi-Square analyses (because the sample sizes are different, Field 2005) reveals that this difference is significant: $\chi^2(1, N = 44) = 8.7, p < .01$.

2.3.4 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 subjects 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 subjects to include more yes responses later in the experiment.

2.4 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 subject 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 requires a non-categorization task for

measuring acceptability, and magnitude estimation fits the bill (Stevens, 1957; Bard et al., 1996).

2.4.1 The benefits of magnitude estimation

Magnitude estimation has been a standard technique in the psychophysical literature for the past 50 years (Stevens, 1957), and in the sociological literature for the past 25 (Lodge, 1981). As briefly outlined in chapter 1, magnitude estimation is simply the measurement (or estimation) of one property of a stimulus using the very same property of a second stimulus as the unit of measure. For instance, the length of the second line below can be estimated using the length of the reference line as a unit of measure. If the reference line is assigned a length value of 100, then second line's length can be estimated using the value 100 for each unit of the reference length: Most

Figure 2.2. Magnitude estimation with line length

Reference: _____
Length: 100

Item: _____
Length: 200

subjects would agree that the second line is twice as long as the reference line, and they would report this by responding 200. Extending this to acceptability judgment collection is straightforward (Bard et al., 1996). In this version of the task, subjects

Figure 2.3. Magnitude estimation with acceptability of sentences

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

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

are asked to rate the acceptability of the second sentence with respect to the reference

sentence. In this case, the acceptability of the mildly unacceptable reference sentence (it is a Whether Island)⁵ is given a numeric value of 100. If the subject thinks that the acceptability of the second sentence is only half that of the reference, they would give it a judgment score of 50. If they think it is 3 times as acceptable, they would give it a judgment score of 300. In this case, the second sentence is a Relative Clause Island, which are generally considered worse than Whether Islands, so one would expect the score to be less than 100.

There are two major benefits of magnitude estimation for our investigation of judgment instability and satiation. First, magnitude estimation eliminates the data loss associated with categorization tasks such as the yes/no task and scale tasks by eliminating all categorization. Because the positive number line is theoretically infinite, subjects may use any number, including fractions and decimals, to indicate a contrast in acceptability no matter how small the difference. Second, because the reference sentence, and thus the unit of measure, remains constant across all items, the numerical distance between two judgments is meaningful. This can be contrasted with the numbers generated by scale tasks, in which the distance between 3 and 4 may differ than the distance between 5 and 6 depending on the categorization strategy of the subject. In statistics, data in which the distances are meaningful are called interval data, whereas data in which the distances are not meaningful are called ordinal data. This distinction is critical, as interval data is required to use the standard battery of parametric inferential statistics (ANOVA, t-test, linear regression, etc.). Ordinal data is amenable to a different set of tests, collectively known as non-parametric tests. For our purposes, the non-categorical, interval level data returned

⁵Mildly unacceptable violations such as the Whether Island are standardly used as reference sentences in linguistic magnitude estimation. The logic behind this is that a mildly unacceptable violation approximates the center of the acceptability scale, thus making it trivial to design an experiment in which half of the sentences are more acceptable, and half are less acceptable, than the reference

by magnitude estimation will allow us to run parametric statistical tests to investigate the nature of any instability we may find (for instance, to test whether model 1 or model 2 is more appropriate).

2.4.2 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):⁶

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

⁶There 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.

2.4.2.1 Subjects

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 subjects 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.

2.4.2.2 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 unacceptable and 2:1 distracters to experimental items.⁷

⁷One 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 subjects 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).

Table 2.12. 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

Table 2.13. Exemplars of each of the Islands investigated

Subject	What do you think a movie about would be scary?
Adjunct	What 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?

Both the experimental items and distracters 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 known acceptability differences with D-linked *wh*-phrases (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 variety (crucially including *who* and *what* to avoid response strategies).

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 subjects 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 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 (ie. 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:

- (1) 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 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 has been claimed to satiate.
 - iv. Subject and Adjunct Islands were administered over the internet, the others in the lab.

2.4.2.3 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 transformation options available, 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 regression, 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 2.14. Linear regressions for means of magnitude estimation in a balanced design

Island Violation	b (y-intercept)	Exp(b) (slope)	p value
Subject Island	-0.13	0.003	.14
Adjunct Island	-0.32	0.003	.52
Whether Island	0.08	0.008	.14
CNPC Island	0.001	-0.006	.44
CNPC with context	-0.02	0.01	.22

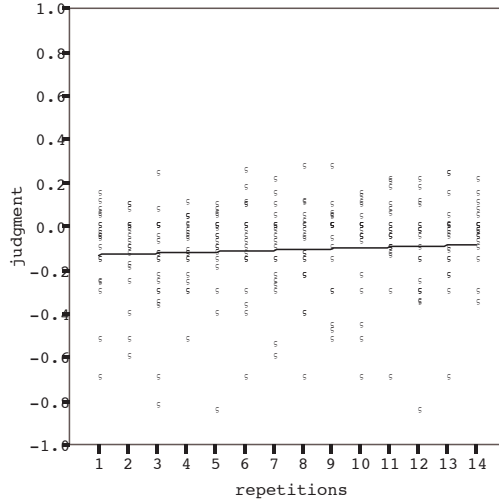
Of course, as demonstrated by model 2 above, it is possible that the satiation effect is found in the variation or spread of the scores over time, not in their means. Therefore a repeated measures linear regression was also performed on the residual scores. Residual scores are the absolute value of the 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.

Table 2.15. Linear regressions for residuals of magnitude estimation in a balanced design

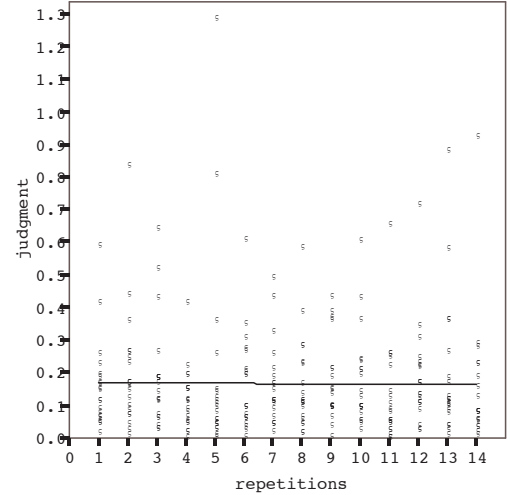
Island Violation	b (y-intercept)	Exp(b) (slope)	p value
Subject Island	0.17	-0.0004	.83
Adjunct Island	0.26	0.004	.26
Whether Island	0.26	-0.002	.68
CNPC Island	0.27	0.01	.18
CNPC with context	0.35	-0.007	.25

Figure 2.4. Scatterplots and trendlines for Subject Islands

Means



Residuals



2.4.2.4 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 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.

2.4.2.5 The memory confound

As an anonymous reviewer correctly points out, the stability we've seen under balanced designs could be due to a confound that is always present in balanced designs: balanced designs increase the likelihood that subjects are able to track their responses, and this memorization may lead to the statistical stability. While the instructions explicitly direct subjects 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 subjects 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 subjects reported noticing such similarities, however, given that the subjects had no training in linguistics, they may have lacked the vocabulary to describe their intuitions. Indeed, many of the subjects were highly cognizant of their 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 subject. Of course, such consistency could be genuine (the subject could be very good at characterizing acceptability), but to be conservative, a second set of linear regressions were performed after REMOVING any subject that reported the same judgment 5 or more times. Given the nature of memorization, one might choose to only remove subjects with repeated scores later in the course of the experiment, or to only remove subjects with consecutive identical judgments. However, to be safe, subjects with repeated scores at any point in the experiment, even if none of the repetitions were consecutive, were eliminated. Thus the subjects that remained in the analysis were those with no overt indication of having memorized their judgments.

Linear regressions for the remaining subjects follow. The additional column S/N indicates the number of subjects who showed significant internal satiation in that their individual judgments showed an increase over time (S), and the number of subjects left in the sample (N) after removing subjects that potentially memorized their judgments:

Table 2.16. Linear regressions for means after removing subjects

Island Violation	b (y-intercept)	Exp(b) (slope)	p value	S/N
Subject Island	-0.15	0.005	.11	0/15
Adjunct Island	-0.27	0.006	.41	2/12
Whether Island	0.03	0.01	.19	3/13
CNPC Island	0.033	-0.004	.67	1/9
CNPC with context	0.032	0.005	.65	2/13 ⁸

As can be seen, judgments were statistically stable even after removing the obviously consistent subjects because they could have potentially memorized their judgments. In fact, even looking at each subject individually yields 3 or fewer satiating subjects 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.

2.4.3 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 subjects 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 satiation (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, 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 isn't clear whether subjects 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.

2.4.3.1 Subjects

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. Subjects also participated in an unrelated self-paced reading experiment during their visit to the lab. All subjects were paid for their participation.

2.4.3.2 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 2.17. 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.

2.4.3.3 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 2.18. Linear regressions for means, CSC-reference

Island Violation	b (y-intercept)	Exp(b) (slope)	p
Adjunct Island	-0.13	-0.0004	.97
CNPC Island	-0.03	0.005	.73
CSC violation	0.02	-0.004	.69
Infinitival Sentential Subject Island	-0.35	0.02	.07
LBC violation	-0.1	-0.003	.84
Relative Clause Island	-0.09	-0.005	.78
Sentential Subject Island	-0.19	0.03	.14
Whether Island	-0.03	-0.004	.79

Table 2.19. Linear regressions for means, If-reference

Island Violation	b (y-intercept)	Exp(b) (slope)	p
Adjunct Island	-0.24	0.01	.64
CNPC Island	-0.14	-0.007	.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	-0.4	0.02	.28
Whether Island	-0.16	0.02	.05

Repeated measures linear regressions were also performed on the residuals:

2.4.3.4 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. 1. However, these three effects were not replicated with the If-reference, as there were no significant effects of variance. There were, however, 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 2.20. Linear regressions for residuals, CSC-reference

Island Violation	b (y-intercept)	Exp(b) (slope)	p
Adjunct Island	0.23	0.003	.63
CNPC Island	0.18	0.005	.61
CSC violation	0.04	0.012	.12
Infinitival Sentential Subject Island	0.32	0.0009	.93
LBC violation	0.17	0.02	.02
Relative Clause Island	0.15	0.03	.02
Sentential Subject Island	0.2	0.006	.72
Whether Island	0.11	0.03	.02

Table 2.21. Linear regressions for residuals, If-reference

Island Violation	b (y-intercept)	Exp(b) (slope)	p
Adjunct Island	0.26	-0.01	.45
CNPC Island	0.18	0.01	.30
CSC violation	0.4	-0.03	.12
Infinitival Sentential Subject Island	0.32	-0.01	.32
LBC violation	0.3	-0.01	.23
Relative Clause Island	0.27	-0.005	.81
Sentential Subject Island	0.32	-0.002	.90
Whether Island	0.17	-0.00007	.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 an alpha level (target p-value) of .05, 20 analyses will nearly guarantee a significant result. A conservative method for correcting this problem is called the Bonferroni correction, and simply involves dividing the alpha level by the number of comparisons being made. Taking the smallest number of comparisons possible, 4, such that the alpha level is the highest possible, it is still only .0125. The only significant effect that achieves this level is in the mean of Infinitival Sentential Subject Islands with the If-reference. Of course, one could still

choose to tell a satiation story for the Infinitival Sentential Subject Island, but such a story would be of limited value given: i) that it has never been tested before, and ii) it is the lowest rated violation in this set. It requires no stretch of the imagination to envision this effect arising from subjects' requiring several repetitions to realize the intended meaning of such an unacceptable violation.

As one anonymous reviewer points out, there may be more going on in the If-reference experiment given that If Islands are structurally identical to Whether Islands. For instance, it could be the case that this structural equivalence means that the subjects 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 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 subjects. 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 doesn't 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 isn't overwhelming evidence for a satiation effect.

2.5 Conclusion

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

Table 2.22. 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 subjects (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 doesn't 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 grammar-based and processing-based violations, 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.

2.6 Some remaining questions

The conclusion that instability can be avoided through balanced designs is not to say that there aren't 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.

2.6.1 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 paper: 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 subjects can easily identify the source of the violation, their judgments may become 'anchored' in a way that isn't 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 subjects and no error terms reported), the order of relative acceptability from highest to lowest was:

Table 2.23. 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 (*want* or *that*), the relative acceptability order corresponds almost directly with the satiation results:

Table 2.24. 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 subjects after the Snyder replications reported in section 2, six subjects 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 correctness on acceptability judgments.

2.6.2 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.

However, as briefly mentioned at the beginning of the paper, one of the replications in this paper was run with an additional task: subjects 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 subjects' confidence about their judgments could be correlated with the factors we have seen influence stability, namely choice of task and choice of design.

Bibliography

- Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72:32–68.
- 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.
- Field, Andy. 2005. *Discovering statistics using spss*. Sage.
- Goodall, Grant. 2005. Satiation and inversion in wh-questions. Talk given at University of Hawaii.
- Hiramatsu, Kazuko. 2000. Accessing linguistic competence: Evidence from children’s and adults’ acceptability judgments. Doctoral Dissertation, University of Connecticut.
- Kaan, Edith, and Laurie Stowe. ms. *Developing an experiment*.
- Keller, Frank. 2000. Gradience in grammar: Experimental and computational aspects of degree of grammaticality. Doctoral Dissertation, University of Edinburgh.
- 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.
- 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 Cognition* 16:149–157.
- 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.
- Schütze, Carson. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. The University of Chicago Press.
- Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31:575–582.
- Stevens, Stanley Smith. 1957. On the psychophysical law. *Psychological Review* 64:153–181.