A test of the relation between working memory capacity and syntactic island effects

Jon Sprouse
Department of Cognitive Sciences
University of California, Irvine

Matt Wagers
Department of Linguistics
University of California, Santa Cruz

Colin Phillips
Department of Linguistics
University of Maryland, College Park

Abstract:

The source of syntactic island effects has been a topic of considerable debate within linguistics and psycholinguistics. Explanations fall into three basic categories: grammatical theories, which posit specific grammatical constraints that exclude extraction from islands; grounded theories, which posit grammaticized constraints that have arisen to adapt to constraints on learning or parsing; and reductionist theories, which seek to reduce island effects to non-grammatical constraints on the human sentence parser, such as limited processing resource capacity. In this paper, we present two large-scale studies designed to test the fundamental prediction of one of the most prominent reductionist theories: that island effects should vary as a function of processing resource capacity. We tested over 300 native speakers of English on four different island effect types (Whether, Complex NP, Subject, and Adjunct) using two different acceptability rating tasks (7-point scale and magnitude estimation) and two different measures of working memory capacity (serial recall and *n*-back). We find no evidence of a relationship between working memory capacity and island effects using a variety of linear regression techniques including bootstrap resampling simulations: individual differences in memory capacity tended to account for only 0% to 2% of the variance in the strength of island effects. These results suggest that island effects are more likely to be due to grammatical constraints or grounded grammaticized constraints than limited processing resource capacity.

1. Introduction

Long-distance dependencies in many of the world's languages exhibit an interesting combination of properties: on the one hand, they are unconstrained with respect to length as measured in both number of words and number of clauses (1), but on the other hand, the types of structures that can contain the gap position (2) are limited:

(1) a. What does Susan think that John bought __?
b. What does Sarah believe that Susan thinks that John bought __?
c. What does Bill claim that Sarah believes that Susan thinks that John bought __?

(2) a. *What do you wonder [whether John bought __]?
b. *What did you make [the claim that John bought __]?
c. *What do you think [the speech about __] interrupted the TV show?
d. *What do you worry [if John buys __]?
e. *What did you meet [the scientist who invented __]?
f. *What did [that John wrote __] offend the editor?
g. *What did John buy [a shirt and __]?
h. *Which did John borrow [__ book]?

In this article, we will focus exclusively on wh-dependencies in English, although it should be noted that this pattern holds for many types of long-distance dependencies, such as relativization, topicalization, and adjective-though fronting (3):

(3) a. *I like the car that you wonder [whether John bought __]?
b. *I know who bought most of these cars, but that car, I wonder [whether John bought __]?
c. *Smart though I wonder [whether John is __], I trust him to do simple math.

Following terminology introduced by Ross (1967), we will refer to the bracketed structures in (2) as *islands*, and the unacceptability that arises when the dependency spans an island (i.e., the gap position is inside the island and the displaced element is outside the island) as an *island effect*. Over the past 40 years, a relatively long list of island structures have been identified within the linguistics literature: wh-islands (2a), complex NP islands (2b), subject islands (2c), adjunct islands (2d), relative clause islands (2e), sentential subject islands (2f), coordinate structures (2g), left-branch extractions (2h), factive islands, and negative islands (for review see Szabolcsi and den Dikken 2006). Given the quantity of dependency types and structures that give rise to island effects, it is perhaps unsurprising that there have been several qualitatively different analyses of the source of island effects, each with their own consequences for the architecture of the language faculty. For expository convenience we will classify these analyses into three broad groups: grammatical theories, reductionist theories, and grounded theories.

Grammatical theories are by far the most well-represented group within the linguistics literature. These theories posit grammatical constraints on wh-dependencies, such that island effects are consequences of violating those constraints. The most popular approach is to posit syntactic constraints, sometimes called *island constraints*, on the formation of wh-dependencies, such as Ross's classic Complex NP Constraint (Ross 1967), Chomsky's Subjacency Condition

(Chomsky 1973, 1986), and Huang's Condition on Extraction Domains (Huang 1982). There have also been several influential non-syntactic grammatical constraints, such as the semantic approach to so-called *weak islands* by Szabolcsi and Zwarts (1993), the semantic approach to adjunct islands by Truswell (2007), and the pragmatic approach by Erteschik-Shir (1973) (see also Goldberg 2007). Although the details of the grammatical theories differ significantly from one another, they all share the common assumption that island effects arise due to the grammatical knowledge of the speaker. This assumption has far-reaching consequences for the architecture of the language faculty, as these grammatical constraints provide a classic motivation for abstract, complex theories of grammar. Furthermore, given the relative infrequency of multi-clausal wh-dependencies even in adult-directed speech, island effects raise difficult questions about how children could use their limited input to arrive at a grammar that includes unbounded wh-dependencies that are nonetheless constrained by specific structural configurations. In this way, island effects are also a classic motivation for theories that assume domain-specific constraints on language acquisition

Given the far-reaching consequences of the assumption that island effects reflect grammatical constraints, it is perhaps unsurprising that there is a second class of theories – which we term *reductionist* theories – that explicitly rejects this assumption. Like grammatical theories, reductionist theories vary in their details; however, the unifying assumption of reductionist theories is that island effects can be reduced to independently motivated constraints on the functioning of the human sentence processor, such as limited attention span (Deane 1991), or limited processing resource capacity (Kluender and Kutas 1993, Kluender 1998, Kluender 2004, Hofmeister and Sag 2010). Under reductionist theories, speakers do not internalize structural constraints on the formation of wh-dependencies (i.e., all wh-dependencies are grammatically well formed). Instead, the perception of unacceptability arises as a by-product of the processing requirements of the sentence. It is easy to see why such an approach is theoretically attractive: under certain assumptions, it raises the possibility of simplifying the grammatical theories in a way that also simplifies the learnability problem faced by children.[1] Given the centrality of debates about representational complexity and domain-specificity in linguistics and in cognitive science more broadly, it is important to seriously investigate the consequences of grammatical and reductionist theories of island effects.

The third approach to island effects – grounded theories – represents a middle-ground between grammatical and reductionist theories: grounded theories share with grammatical theories the assumption that the immediate cause of island effects is formal grammatical constraints in the minds of speakers, and they share with reductionist theories the assumption that island effects are ultimately caused by independently motivated properties of the human sentence parser. Grounded theories argue that island effects could have arisen historically because of parsing efficiency considerations, but that this efficiency consideration was ultimately

[1] The purported learnability benefit of reductionist theories is not as straightforward as is often assumed. First, the learnability problem is likely smaller than assumed: some of the island constraints are considered universal by generative grammarians such that only island constraints that vary cross-linguistically need to be learned. Second, and perhaps more importantly, shifting the burden from the theory of grammar to a theory of processing costs means that the learner must identify which distributional facts arise due to grammatical constraints and which distributional facts arise due to processing costs. This adds a layer of complexity to the problem of identifying generalizations from the input.

grammaticized as a set of structural constraints (Fodor 1978, 1983, Berwick and Weinberg 1984, Hawkins 1999). We will have relatively little to say about grounded theories in this article, as they appear to make the same (synchronic) predictions as grammatical theories.

The goal of this article is to make some empirical headway in the debate between grammatical theories and reductionist theories by focusing on a prominent reductionist approach – the capacity-based theory first proposed by Kluender and Kutas 1993 (and expanded in Kluender 1998, 2004, and Hofmeister and Sag 2010). Although several researchers have advocated the capacity-based reductionist approach, the mechanisms and predictions of that theory have not often been spelled out in detail. Therefore we will first discuss the crucial distinction between the capacity-based theory and grammatical theories. Second, we will present the results of two large-scale studies designed to test those predictions directly, and discuss their consequences for capacity-based reductionist theories. To that end, the rest of this article is organized as follows. In Section 2 we discuss the capacity-based theory in detail in an attempt to identify divergent predictions of the capacity-based theory and grammatical theories. We argue that divergent predictions arise only when island effects are defined as a statistical interaction of the acceptability of four conditions. This definition of island effects contrasts with standard definitions within both the grammatical and reductionist traditions, which tend to define island effects in terms of comparisons of pairs of sentences. In Section 3 we outline the logic of the two large-scale studies (with 142 and 173 participants, respectively) that we designed to test for a relationship between processing resource capacity and individual differences in the strength of island effects for different island types, as predicted by the capacity-based theory. Sections 4 and 5 present the results of those two studies. Because neither study reveals any evidence of a relationship between processing resource capacity and island effects, in Section 6 we present a bootstrap-based simulation to determine whether the lack of significant relationship found in the linear regressions in Sections 4 and 5 could have been due to the averaging procedure required by those analyses. Again, the results suggest that there is no evidence of a relationship between resource capacity and island effects, which we interpret as inconsistent with the capacity-based reductionist theories, but compatible with grounded or grammatical approaches. Finally, in Section 7 we discuss several potential objections to our interpretation of this data, and argue that these objections rely on assumptions about working memory that are extremely unlikely to be true. We conclude that the most profitable avenue for future research into the nature of island effects is either the grounded approach, which maintains many of the attractive properties of the reductionist approach (but treats the effect of capacity constraints in evolutionary terms, rather than synchronic effects), or the grammatical approach, which is actively being researched by many linguists.

2. The capacity-based theory and the definition of island effects

The assumptions of the capacity-based theory

We begin with the capacity-based theory of island effects proposed by Kluender and Kutas (1993), which argues that island effects arise because the parser lacks the processing resources necessary to successfully parse the critical sentences. The central observation of the capacity-based theory is that the sentences that give rise to island effects always contain two specific components: (i) a long-distance (often bi-clausal) wh-dependency, and (ii) a complex syntactic structure (which we call *island* structures). Under the assumption that limited resources are

available for the parsing processes necessary in any given sentence, these two components (dependency length and island structure) might together lead to an overload of processing resource capacity. Formalizing this analysis requires the following assumptions:

(4) Assumptions of the capacity-based theory (Kluender and Kutas 1993)

i. There is a limited pool of processing resources available that must be shared by all simultaneous processes

ii. Unacceptability arises if the simultaneous processes necessary to complete the parse require more resources than are available

iii. There is a processing cost associated with the operations necessary to build long-distance wh-dependencies

iv. There is a processing cost associated with the operations necessary to build the syntactic structures that we call *island* structures

v. These two (sets of) processes must be deployed simultaneously in sentences that give rise to island effects

These five assumptions entail that the sentences that give rise to island effects have (at least) two sets of resource intensive processes that must be deployed simultaneously: the processes necessary to construct the long-distance wh-dependency and the processes necessary to build the island structure. Of course, some care is necessary in defining "simultaneously" in these assumptions: it is not the case that island effects arise whenever an island structure and an incomplete wh-dependency occur simultaneously. For example, (5a) below contains a wh-dependency that spans a relative clause in subject position (a "double" island structure) and is considered grammatical according to the informal judgments of linguists. The island effect only arises when the wh-dependency terminates within the island structure as in (5b).

(5) a. Who did [the reporter that won the Pulitzer prize] interview __ for the expose?
b. *Who did [the reporter that interviewed __] win the Pulitzer prize?

To our knowledge the complexity of the definition of "simultaneous" in the capacity-based theory has not been addressed in the literature. This is a potential problem; however, for the sake of argument we will simply assume that a suitable formulation could be found such that the crucial capacity overload only occurs when the wh-dependency terminates within the island structure. If one grants the assumptions in (4), and grants that the simultaneity problem in (5) can be solved, then island effects can be reduced to a consequence of limited resource capacity rather than a consequence of formal grammatical constraints. To the extent that each of these assumptions is needed independently of the existence of island effects, the capacity-based theory reduces the number of assumptions necessary to explain island effects as compared to grammatical theories, which always require the additional assumption of an explicit grammatical constraint on the formation of long-distance dependencies.

It is, of course, an empirical question whether each of the assumptions in (4) is needed independently of the existence of island effects, and there is no clear consensus in the field regarding these issues. Kluender and Kutas (1993) argue that there are ERP correlates of each of the two processing costs (assumptions (iii) and (iv), and that the other assumptions, such as limited processing resource capacity, are shared by many (but not all) within the sentence processing literature (e.g., King and Just 1991, Just and Carpenter 1992, Caplan and Waters 1999, Fedorenko et al 2006, 2007). In this way, Kluender and Kutas (1993) provide a proof-of-concept argument for the capacity-based theory (at least for *if* and *wh*-islands; see Kluender 1998 for relative clause islands, and Kluender 2004 for subject islands). In this article, we focus primarily on the constrained capacity assumption itself rather than the individual processing costs associated with dependency length and island structures. However, it should be noted that the experiments presented here do contain evidence about the robustness of the effect of each of these potential processing costs on acceptability judgments: the island-structure cost is only reliably present in the acceptability judgments of whether islands, and is reliably absent in complex NP islands and subject islands. In contrast, the effect of dependency length is relatively robust. The unreliability of the island-structure cost raises a potential problem for the capacity-based theory, as it is not clear how island effects could arise from a single processing cost. The reliability of the island-structure cost is discussed briefly in the results sections of each of the experiments. However, for space reasons, in the discussion sections we have chosen to focus primarily on the core assumption that limited capacity is the source of island effects.

The definition of island effect

At this point it should be clear that the capacity-based theory defines island effects as a *psychological interaction* of two (sets of) parsing processes that occurs because the processes rely upon a single pool of resources. As noted by Kluender and Kutas (1993), for purposes of acceptability judgment experiments this psychological interaction can be translated into a *statistical interaction* between two factors, each with two levels: LENGTH (short, long) and STRUCTURE (non-island, island). The factor LENGTH manipulates the length of the wh-dependency at two levels: within a bi-clausal constituent question, a short dependency is created by extraction of the matrix subject; a long dependency is created by extraction of the object argument in the embedded clause. The factor STRUCTURE refers to the STRUCTURE of the embedded clause.

Table 1: Independent manipulation of dependency length and island structures

LENGTH	STRUCTURE	Example
short	non-island	Who __ thinks that John bought a car?
long	non-island	What do you think that John bought __?
short	island	Who __ wonders whether John bought a car?
long	island	What do you wonder whether John bought __?

Defining island effects in this way has several advantages. First, it allows us to isolate the effect of each of the individual factors on continuous acceptability ratings. For example, the effect of processing long-distance wh-dependencies can be seen by comparing the short, non-island condition to the long, non-island condition, and the effect of processing island structures can be

seen by comparing the short, non-island condition to the short, island condition. Second, it allows us to quantify the statistical interaction of the two factors. If there were no statistical interaction between the two factors (i.e., if the two sets of processes impact acceptability ratings independently), we would expect a graph like that in Figure 1a. Figure 1a is an example of simple linear additivity between each factor in which the cost of each process leads to a decrement in acceptability ratings, and in which each cost sums linearly with respect to the short/non-island condition. This linear additivity in decrements leads to two parallel lines. However, if there were an interaction between the two factors, we would expect a graph like that in Figure 1b: super-additivity when the *long* and *island* levels of the two factors are combined, leading to non-parallel lines. (The hypothetical ratings in Figure 1 are displayed in terms of standardized z-scores, which can be derived from any approximately continuous rating measure, such as Likert scales or magnitude estimation.)

Figure 1: Example results for main effects and interaction

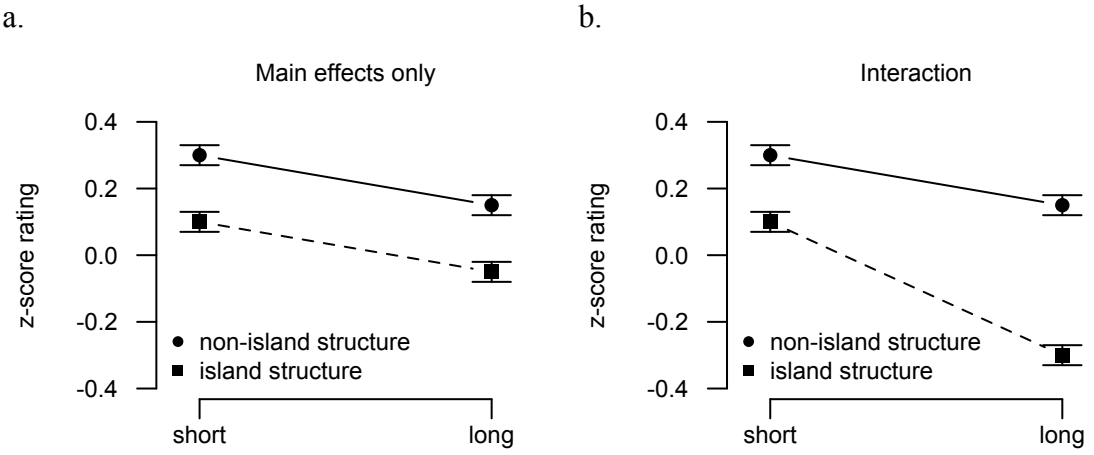


Figure 1b is in fact the pattern that is consistently observed when the factors LENGTH and STRUCTURE are independently manipulated in acceptability experiments, although there is variation in the size of the effect of island structures alone (Kluender and Kutas 1993, Sprouse 2007, Sprouse et al. 2010). In this way, island effects can be defined as a statistical interaction between two structural factors, exemplified by a super-additive decrease in acceptability in the long, island condition.

Distinguishing the capacity-based theory and grammatical theories

Having defined island effects quantitatively, in terms of a superadditive effect on acceptability ratings when long dependencies are combined with island structures, it is now possible to articulate a fundamental difference between the capacity-based theory and grammatical theories. The pattern that we obtain by using the quantitative definition of island effects shows a *statistical interaction* of the two experimental factors. Capacity-based theories interpret this statistical interaction as reflecting a *psychological interaction* of the two sets of processes represented by the two experimental factors, due to a shared pool of limited resources. Grammatical theories, on the other hand, do not interpret the statistical interaction as a psychological interaction at all.

Instead, grammatical theories attribute the extreme unacceptability of the long, island condition as resulting from a grammatical constraint that impacts only sentences that contain long-distance dependencies that enter island structures. Crucially, then, the two theories differ in their account of the psychological cause of the statistical interaction: the capacity-based theory attributes the statistical interaction to a limited pool of processing resources, and the grammatical theories attribute the statistical interaction to a grammatical constraint that targets the long, island condition.

The capacity-based and grammatical theories do not necessarily differ in their predictions about the acceptability of individual conditions, or even the relative acceptability of pairs of conditions. These two theories only differ when it comes to the source of the interaction, as the two theories ascribe different underlying causes to the interaction. By manipulating factors that affect those underlying causes, we should be able to observe correlated changes in the statistical interaction. Under the capacity-based theory, the strength of the statistical interaction in acceptability judgments should co-vary with the availability of processing resources, as it is the limited quantity of processing resources that causes the interaction. Specifically, under the capacity-based theory individuals with larger processing resource capacity should exhibit weaker statistical interactions, and individuals with more limited processing resource capacity should demonstrate stronger statistical interactions. Under the grammatical theory, processing resource capacity plays no role in the statistical interaction, therefore any individual differences in the strength of island effects that we might see in the population should not correlate with individual differences in processing resource capacity.

3 The logic of the present studies

In order to investigate this prediction of the capacity-based theory, we need: (i) a measure of processing resource capacity, and (ii) a measure of the strength of the interaction that we have used to define island effects. In this section, we discuss the rationale for the measures that we chose, as well as the specific statistical predictions of the capacity-based and grammatical theories.

Working memory as a measure of processing resource capacity

There are a number of measures that might reflect the capacity (or flexibility) of an individual's processing resources, and there is evidence that many of these measures correlate with real-time sentence-processing performance in reaction times and error rates (King & Just, 1991, Just & Carpenter, 1992, MacDonald, Just, & Carpenter 1992, Caplan and Waters 1999, Vos et al. 2001, but cf. Roberts and Gibson 2002). For the present studies, we chose to use both the serial recall task (Experiments 1 and 2) and the *n*-back task (Experiment 2) to derive capacity measures for each participant. In the serial recall task participants are presented with a series of words one at a time, and when the presentation is complete, they are asked to recall those words in the order that they were presented. In the *n*-back task, participants are presented with a series of letters on a computer screen one at a time (RSVP), and are asked to press a button if the letter currently on the screen was also presented *n* items previously (Kirchner 1958, Kane and Engle 2002, Jaeggi et al 2008). This means that in order to successfully complete the task, the participant must continuously update the *n* letters that are kept in memory through the entire presentation (in our experiments, 30 letters were presented in sequence during each trial). By increasing the value of

n (in our experiments, participants completed a 2-back, 3-back, and a 4-back task, in that order), the experimenter can increase the difficulty of the task to obtain a working memory capacity measure.

Our rationale for these choices was as follows. First, we wanted to follow the suggestion by Kluender and Kutas (1993) that verbal working memory capacity is a likely common resource required by the two sets of processes (see also Kluender 1998, 2004, and Ueno and Kluender 2009). Second, we wanted to avoid any potential confound between the rating task and the memory task, which meant avoiding sentence-based memory tasks (e.g., the reading span task of Daneman and Carpenter 1980). Finally, we wanted to maximize the possibility of finding a significant correlation between working memory capacity and island effects. Most working memory capacity measures appear to be highly correlated with one another (Conway et al. 2005), and in fact, in a confirmatory factor analysis, immediate free recall was found to be as strong a measure of working memory capacity as complex span tasks (Unsworth and Engle 2007). Furthermore, the *n*-back task has been shown to be uncorrelated with standard memory span tasks (Kane et al. 2007), which suggests that including both serial recall and *n*-back tasks maximizes the possibility of tapping into the appropriate component of processing capacity that varies between individuals.

Measuring the strength of island effects

As discussed in Section 2, the crucial difference between capacity-based and grammatical theories involves the source of the statistical interaction elicited by designs like Table 1. Therefore we need a measure that captures the strength of the statistical interaction, and can be compared to the working memory measures discussed above. A standard measure known as a *differences-in-differences* (DD) score achieves just this (Maxwell and Delaney 2003). A DD score is calculated for a two-way interaction as follows: First, calculate the difference (D1) between two of the four conditions. To make the DD scores as intuitively meaningful as possible, we have defined D1 as the difference between the long, non-island rating and the long, island rating. Second, calculate the difference (D2) between the other two conditions. For our purposes, D2 is the difference between the short, non-island rating and the short, island rating. Finally, calculate the difference between these two difference scores:

(6) **Calculating the DD score with a sample set of mean ratings**

a. D1 = (long, non-island) – (long, island) — rating (*z*-score units)

What do you think that John bought ___?	0.5
What do you wonder whether John bought __?	– -1.5
	2.0

b. D2 = (short, non-island) – (short, island)

Who __ thinks that John bought a car?	1.5
Who __ wonders whether John bought a car?	– 0.7
	0.8

c. DD = D1 – D2 = 2.0 – 0.8 = 1.2

Because DD scores can be calculated for each individual tested (using standard continuous acceptability judgment experiments), DD scores can serve as a composite measure of the strength of the statistical interaction for each individual and intuitively can be thought of as the strength of the island effect for that individual: a positive DD score reflects a super-additive interaction, with larger values representing larger interactions (stronger island effects); a DD score of 0 represents no interaction at all (no island effect).

Statistical predictions of the capacity-based and grammatical theories

The capacity-based theory makes the prediction that if an individual has sufficient working memory capacity to handle both sets of processes simultaneously, then that individual's DD score (island strength) might be 0. However, there are obvious constraints on the amount of working memory capacity that any individual can have, so it is an empirical question whether this theoretical limit could ever be reached. There is some reason to believe that proponents of the capacity-based theory believe that sufficient working memory capacity is humanly possible. For example, Hofmeister and Sag (2010) make the following informal observations as potential support for the capacity-based theory:

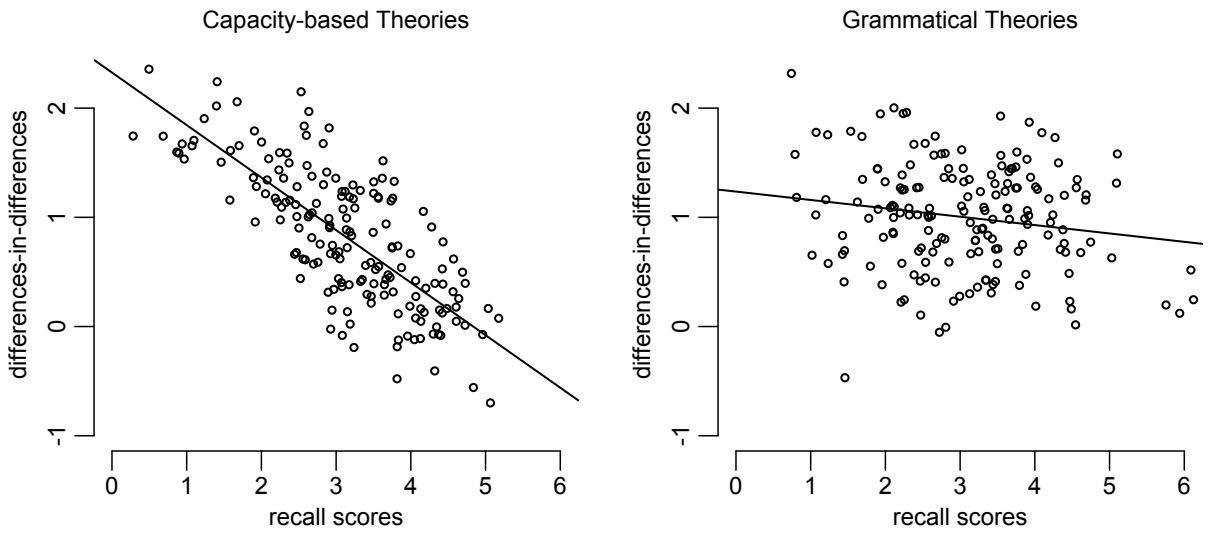
> Notably, some individuals seem fairly accepting of island violations, while others reject the same tokens. This type of variation in acceptability judgments, both within and across subjects emerges naturally on the processing account of islands. Individuals are known to differ significantly from one another in terms of working memory capacity (Daneman & Carpenter 1980; King & Just 1991; Just & Carpenter 1992), and the same individual may have more or fewer resources available, depending upon factors such as fatigue, distractions, or other concurrent tasks. [Hofmeister and Sag 2010:56]

In the current studies we test the more general prediction that there should be a significant inverse relationship across individuals between the strength of the island effect (DD scores) and working memory capacity, which may or may not include individuals that report no island effects (in our measures, a DD score of zero).

For example, if we were to plot DD scores as a function of working memory capacity for a sufficiently large sample of speakers, the capacity-based theory predicts that the we should see a downward sloping trend as schematized in Figure 2a: as working memory scores increase, DD scores should decrease. Statistically speaking, the capacity-based theory predicts that working memory capacity should be a significant predictor of DD scores (e.g., using a standard linear regression), such that the line of best fit derived for the relationship should (i) have a negative slope, and (ii) account for a relatively large portion of the variance in the sample, i.e., measures of goodness of fit such as R^2 should be relatively large. On the other hand, grammatical theories predict no relationship between variation in DD scores and variation in working memory scores, as schematized in Figure 2b. Statistically speaking, grammatical theories predict that working memory capacity should not be a significant predictor of DD scores, such that the line of best fit derived for the relationship should not account for much of the variance in the sample at all, i.e., a low R^2 value.

Figure 2: Predictions of the capacity-based and grammatical theories

a. b.



In the sections that follow, we present two large-scale experiments that were designed to test these predictions. Experiment 1 tested 142 undergraduates using the serial recall task and a 7-point acceptability judgment task. Experiment 2 tested a separate group of 173 undergraduates. It expanded Experiment 1 by using both a serial recall task and an *n*-back task for estimating working memory capacity and by using the magnitude estimation task for collecting acceptability judgments. Together, these two experiments tested over 300 individuals using two different types of working memory measures and two different types of acceptability judgment tasks.

4. Experiment 1

In this section, we present the first of our two large-scale studies. The four island types investigated in this experiment (and Experiment 2) were chosen because they are considered to be relatively weak island effects compared to many of the other island types (not to be confused with "weak islands", which is a theoretical distinction between types of islands (Szabolcsi and den Dikken 2006)). In fact, each of the four island types chosen have a relatively stronger counterpart: whether island – wh-island, complex NP island – relative clause island, subject island – sentential subject island, (conditional) adjunct island – causal adjunct island. The weaker versions of these island types are generally considered more interpretable by linguists, and therefore should be the best possible candidates to display the correlation with working memory capacity predicted by reductionist theories.

4.1 Materials and methods

Participants

142 (76 female) self-reported monolingual native speakers of English, all University of California Irvine undergraduates, participated in this experiment for course credit. The experiment was administered during a single visit to the lab. Participants completed both the acceptability rating task and the serial recall task during their visit. The acceptability rating task always preceded the serial recall task.

The acceptability rating task

Four types of island effects were tested using the design described in Section 2, for a total of 16 critical conditions: whether islands, complex NP islands, subject islands, and adjunct islands. Eight additional sentence types were included to add some variety to the materials, for a total of 24 sentence types in the survey. All of the conditions were wh-questions. Eight tokens of each sentence type were created, and distributed among 8 lists using a Latin Square. The 8 lists were then combined in pairs creating 4 master lists containing 2 tokens of each condition procedure, such that related lexicalizations never appeared in the same list. Two pseudorandom orders of each of the 4 lists were created, such that related conditions never appeared in succession. This resulted in 8 lists of 48 items in pseudorandom order, with each list containing 2 tokens of each condition.

Table 2: Example materials for Experiment 1

WHETHER ISLAND CONDITIONS	COMPLEX NP CONSTRAINT CONDITIONS
Who __ thinks that John bought a car?	Who __ claimed that John bought a car?
What do you think that John bought __?	What did you claim that John bought __?
Who __ wonders whether John bought a car?	Who __ made the claim that John bought a car?
What do you wonder whether John bought __?	What did you make the claim that John bought __?
SUBJECT ISLAND CONDITIONS	**ADJUNCT ISLAND CONDITIONS**
What do you think the speech interrupted __?	Who __ thinks that John left his briefcase at the office?
What do you think __ interrupted the TV show?	What do you think that John left __ at the office?
What do you think the speech about global warming interrupted the TV show about __?	Who __ laughs if John leaves his briefcase at the office?
What do you think the speech about __ interrupted the TV show about global warming?	What do you laugh if John leaves __ at the office?

The acceptability rating task was presented as a paper survey. Six practice items were added to the beginning of each survey (two each of low, medium, and high acceptability). These practice items were not marked as such, i.e., the participants did not know they were practice items, and did not vary in order or lexicalization between participants. Including the practice items, the surveys were 54 items long. The task was a standard 7-point scale acceptability judgment task where 1 represented “least acceptable”, and 7 represented “most acceptable”. The directions and

materials are available on the corresponding author's website. Participants were under no time constraints during their visit.

The serial recall task

The serial recall task used 8 disyllabic words that were matched for orthographic and phonetic form (CVCVC), approximate frequency, neighborhood density, and phonotactic probability. The 8 words were: *bagel, humor, level, magic, novel, topic, tulip, woman*. The 8 words were recorded by a female native speaker for auditory presentation to the participants. We created 10 auditory lists, each containing the full set of 8 words in a different order. The same 8 words were used in each list to prevent the use of mnemonics during the memorization stage (Cowan 2000).

Each participant was presented with all 10 sequences in the same order. The words in each list were presented sequentially with an ISI of 500ms. Participants were instructed to repeat the word *the* quietly to themselves during the auditory presentation in order to suppress articulatory repetition of the list during presentation (Cowan 2000). The trials were presented auditorily using a computer and headphones in a private testing room. Participants were given 30 seconds to recall the list following each trial, and were asked to do so using a pen or pencil on a paper scoring sheet, to avoid penalizing the responses of slow or inaccurate typers.

The standard procedure for scoring serial recall tasks is as follows: First, within each trial, a response is counted as correct only if it appears in the correct position in the response list (1-8). Second within each position across trials, the total number of correct responses is summed, and divided by the number of trials (10) to derive the proportion correct (between 0 and 1) for each position. Finally, the proportions correct for all of the positions are summed to derive a memory span score (between 0 and 8) for each participant. Unfortunately, the instructions for the serial recall task in Experiment 1 did not instruct participants to leave blank responses for the positions that they did not remember. This could have had the unintended consequence of leading participants that correctly remembered words 2-8 to write those words in slots 1-7, thus receiving a score of 0. To correct for this, we adopted a slightly different scoring procedure for Experiment 1: a response within a trial was counted as correct if the response that immediately precedes it is the immediately preceding word in the list. This is a slightly stricter scoring procedure than the standard procedure, but it preserves the serial component of the task, and gives the hypothetical participant described above credit for his responses: in this case, the first response would be incorrect because there was no immediately preceding response, but the following six responses would be counted as correct. The instructions were modified in Experiment 2, such that Experiment 2 could adopt the standard (and slightly less strict) scoring procedure.

4.2 Results

For the acceptability judgment task, each participant's ratings were z-score transformed prior to analysis. The z-score transformation is a standardization procedure that corrects for some kinds of scale bias between participants by converting a participant's scores into units that convey the number of standard deviations each score is from that participant's mean score. The z-score transformation eliminates the influence of scale bias on the size of the DD scores, and therefore

increases the likelihood of finding a significant relationship between working memory capacity and DD scores.[2] The means and standard deviations for each condition are reported in Table 3:

Table 3: Experiment 1, means and standard deviations for each condition (n=142)

	whether	complex NP	subject[3]	adjunct
short, non-island	0.87 (0.60)	0.86 (0.58)	0.45 (0.86)	0.77 (0.64)
long, non-island	0.22 (0.76)	0.20 (0.78)	0.74 (0.78)	0.28 (0.81)
short, island	0.47 (0.69)	0.83 (0.61)	-0.43 (0.73)	0.61 (0.65)
long, island	-0.91 (0.60)	-0.81 (0.65)	-0.95 (0.61)	-0.92 (0.63)

The basic island effects

The first question we can ask about this data is whether each of the island effects, as defined in Section 2, are present in this rating study. We ran linear mixed effects models with items and participants included as random factors on each of the island types using LENGTH and STRUCTURE as fixed factors (comparable to a repeated-measures two-way ANOVA, but with participants and items entering the model simultaneously). All *p*-values were estimated using the MCMC method implemented in the languageR package for R (Baayen 2007, Baayen et al 2008). Table 4 reports the *p*-values for each factor and the interaction.

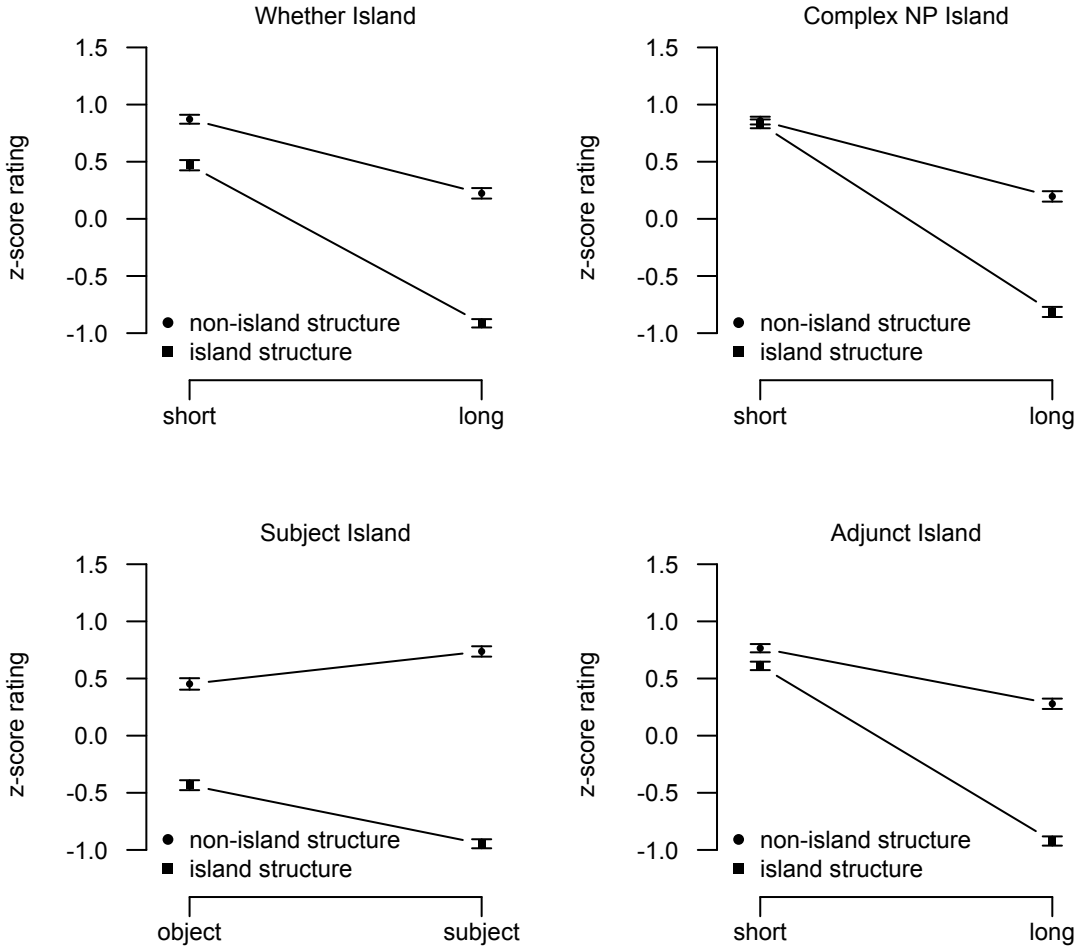
Table 4: Experiment 1, two-way linear mixed effects models for each island type (n=142)

	whether	complex NP	subject	adjunct
LENGTH	.0001	.0001	.0001	.0001
STRUCTURE	.0001	.6156	.0001	.0946
LENGTH x STRUCTURE	.0001	.0001	.0001	.0001

[2] We also ran all of the regression analyses reported for Experiment 1 and Experiment 2 using the raw scores rather than the z-score transformed scores with no change in results. None of the raw score models captured a meaningful portion of the variance (usually between 0-1%).

[3] Note that the subject island sub-design in experiment 1 is slightly different from the rest. The short level of the length factor refers to a gap in the object position, and the long level refers to a gap in the subject position.

Figure 3: Experiment 1, interaction plots for each island type (n=142)

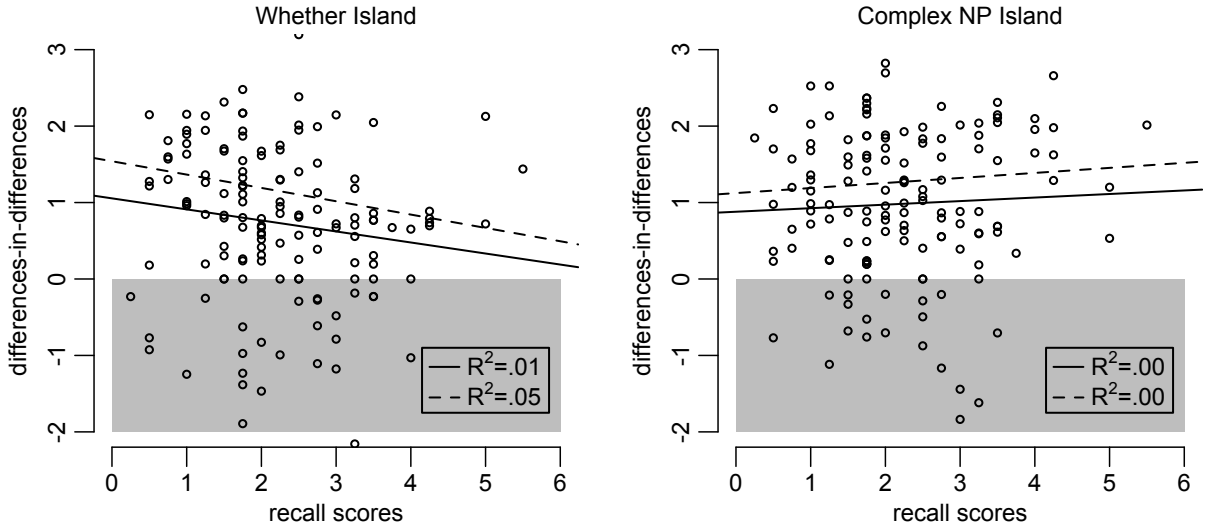


There is a significant main effect of LENGTH for each island type. There is only a significant main effect of STRUCTURE for the whether and subject island types; complex NP and adjunct islands did not show a main effect of STRUCTURE. And, crucially, there is a significant interaction of LENGTH and STRUCTURE for every island type (at $p < .0001$), suggesting that there are indeed island effects for each of these island types. However, the pattern of results for complex NP and adjunct islands is not as predicted by the capacity-based theory: there is a significant island effect (interaction) without any evidence of a cost to the island structure independently (no main effect of STRUCTURE). This pattern of results raises a significant problem for the generalizability of the capacity-based theory, as one of the fundamental processing costs does not appear to be robust in all of the island types (even with our extremely large sample size of 142). This raises the question of how island effects could be the result of a conspiracy of two processing costs when acceptability ratings show evidence of one of the processing costs in only some of the island types. It should also be noted that the relatively large effect of STRUCTURE in subject islands may be an artifact of the slightly different design used for subject islands – a possibility corroborated by the lack of main effect of STRUCTURE for the corrected subject island design used in Experiment 2 (see Section 5).

Differences-in-differences as a function of serial recall

Scores on the recall task ranged from 0.25 to 5.5, with a mean of 2.21 and a standard deviation of 1.03. DD scores were calculated following the formula given in (6) and plotted as a function of serial recall scores in Figure 4. Two sets of simple linear regressions were run for each island type using the serial recall and DD scores. The first set of regressions was run on the complete set of DD scores for each island type. The second set of linear regressions were run on only the DD scores that were greater than zero for each island type. The logic behind the second analysis is that DD scores below 0 are indicative of a *sub-additive* interaction. Neither theory predicts the existence of sub-additivity, which suggests that DD scores below 0 may reflect a type of noise that we may not want to influence the linear regression. By eliminating these potentially unrepresentative scores from the analysis, we increase the likelihood of finding a significant trend in the data. This affected 27 participants for whether islands, 20 participants for complex NP islands, 19 participants for subject islands, and 16 participants for adjunct islands. Table 5 reports the results of the two sets of linear regressions.

Figure 4: Experiment 1, differences-in-differences scores plotted as a function of serial recall scores (n=142). The solid line represents the line of best fit for all of the DD scores. The dashed line represents the line of best fit when DD scores below 0 are removed from the analysis (shaded grey). Trend lines were fitted using a least-squares procedure. Adjusted R^2 for each trend line is reported in the legend.



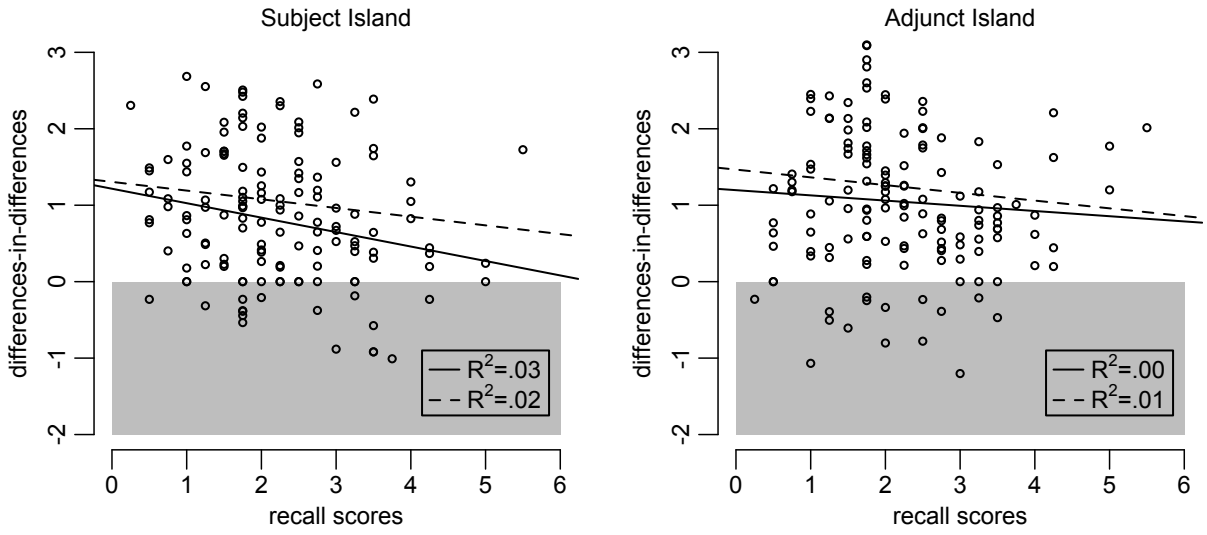


Table 5: Experiment 1, linear regression modeling differences-in-differences scores as a function of serial recall scores (n=142)

	island	intercept	slope of recall	t-statistic of recall	p-value of recall	adjusted R^2
All DDs	whether	1.05	-0.14	-1.65	.10	.01
	complex NP	0.88	0.05	0.58	.56	.00
	subject	1.22	-0.19	-2.42	.02	.03
	adjunct	1.20	-0.07	-0.92	.36	.00
DDs greater than zero	whether	1.54	-0.17	-2.67	.01	.05
	complex NP	1.12	0.07	1.08	.28	.00
	subject	1.31	-0.11	-1.77	.08	.02
	adjunct	1.47	-0.10	-1.56	.12	.01

A simple linear regression analysis finds the line that minimizes the distance between all of the points and the line itself, and reports the coefficients of that line: its intercept with the y-axis and the slope associated with a one unit change in the predictor variable, which in this case corresponds to recall scores. As with all modeling procedures, a line is always returned by the least-squares procedure, so the first question is whether this line explains the data significantly better than other possible lines, such as a line chosen at random, or a horizontal line that uses the mean as a y-intercept. The adjusted R^2 statistic is a direct measure of the goodness of fit of the line: adjusted R^2 describes the proportion of the variance in the data captured by the line (between 0 and 1) with a slight adjustment based on the number of degrees of freedom in the model. As Table 5 reports, 5 of the 8 models accounted for 0% or 1% of the variance in their respective data sets. The best possible model was one of the biased models, and it only captured 5% of the variance in its data set. As a point of comparison, the line of best fit in the graph in Figure 2 that we used to illustrate the prediction of the capacity-based theory has an R^2 of .5 (i.e., 50% of the variance in the data is explained by the line), which within the psycholinguistics literature is generally considered to be a highly meaningful correlation. Unlike *p*-values, there are no broadly agreed-upon conventions for interpreting R^2 values; however, it is safe to assume

that the extremely small R^2 values found for each of the island types (even after removing noisy DD scores) are not at all what one would predict for a theory like the capacity-based theory, which relies heavily on a single factor for its explanatory power. These results strongly suggest that there is no evidence of a relationship between DD scores and recall scores in Experiment 1.

A second potentially interesting statistic in these models is the t-statistic of the slope of the predictor variable. This statistic quantifies the difference between this slope and a slope of 0 (i.e., a horizontal line), and the associated *p*-value reports the likelihood of observing a slope with this value in a sample if the actual population slope is 0. As Table 5 reports, only two of the 8 regressions returned lines with slopes that are significantly different from 0 at a level of $p < .05$, one island type in each of the sets of regressions, and a third line that is marginally significant ($p = .08$). If the models accounted for a significant portion of the variance (high R^2 values), we would then look for significant slopes to confirm that the relationship was indeed non-zero (negative or positive). In this case, however, these models do not explain significantly more variance than would a randomly selected line, so it is difficult to attach any meaningful interpretation to the slope of these two lines. The unreliability of the slopes of poorly fitting trend lines becomes readily apparent when one compares the results of experiment 1 to the results of experiment 2 (section 5.2): in both experiments there are two (out of eight) lines with slopes that are significantly different 0. In experiment 1, both of the lines have a negative slope; in experiment 2, both of the lines have a positive slope. These perfectly conflicting results underscore the fact that the slopes of poorly fitting trend lines are not meaningful.

The results of the linear regressions reported above suggest that there is no relationship between DD scores (a measure of the strength of island effects) and serial recall scores (a measure of working memory capacity) in Experiment 1. However, as with all null results, before we can be confident in the conclusion that there is indeed no relationship, we must be confident that the failure to find an effect was not due to the design of the experiment. For example, one possible objection to the analyses above is that we employed a serial recall scoring metric that is stricter than the standard scoring metric. To control for this, we also employed a metric that is less strict than the standard scoring procedure with nearly identical results. Of course, not all possible concerns can be addressed with a change in analysis. Therefore, in Experiment 2 addressed a number of possible concerns with the design of Experiment 1 that might have contributed to the failure to find a significant relationship. For example, another possible concern with the serial recall task in Experiment 1 involves the task itself: perhaps serial recall does not capture the relevant components of working memory. To minimize the possibility that the null effect found in Experiment 1 was due to the choice of the capacity measure, we included a series of *n*-back tasks in Experiment 2, as the *n*-back task has been recently shown to be relatively distinct from serial recall in the components of working memory that it measures (Kane et al. 2007).

Turning to the acceptability judgment task, one possible concern is that Experiment 1 presented only 2 tokens per condition to each participant, which could have contributed some noise to the DD scores. Therefore, in Experiment 2 we increased the number of tokens per condition to 4 tokens per condition. Another potential concern is that the 4 conditions for the subject island sub-design in Experiment 1 differed from the 4 conditions for the other island types. The design in Experiment 1 is more like the one that appears in the theoretical syntax literature, but crucially leads to a much smaller interaction than the standard STRUCTURE X LENGTH design of the other island types. This could have led to a limited range of variation in the subject island DD scores. Therefore, the subject island sub-design in Experiment 2 was modified

to use the standard STRUCTURE x LENGTH design. It is also possible that the 7-point response scale used in Experiment 1 could have compressed the range of possible ratings, as the 7-point scale imposes a ceiling and floor on the scale. Experiment 2 used the magnitude estimation task (Bard et al. 1996, Keller 2000) in an attempt to eliminate the potential ceiling effects and mitigate the potential floor effects. Finally, the composition of the survey was such that there were more acceptable sentences than unacceptable sentences (a ratio of 5:3). The reason for this was to keep the acceptability judgment survey relatively short (54 items) because the lab visit also involved the serial recall task. However, the asymmetry may have increased the saliency of the ungrammatical sentences, potentially reducing the variation in ratings. The acceptability judgment survey in Experiment 2 maintained a ratio of 1:1 for acceptability, and also balanced the ratio of declarative to interrogative sentences, and the ratio of target sentences to filler sentences.

5. Experiment 2

Experiment 2 tested the same 4 island types as Experiment 1, but used two different measures of working memory capacity, the serial recall task used in Experiment 1 and a series of *n*-back tasks, and a different acceptability rating measure, magnitude estimation. Magnitude estimation is a task in which participants are asked to judge the relative difference between successive test stimuli and an experiment-wide standard stimulus (Stevens, 1956). Participants are presented with a physical stimulus, such as a light source set at a pre-specified brightness by the experimenter. This physical stimulus is known as the *standard*. The standard is paired with a numerical value, which is called the *modulus*. The participants are told that the brightness of the light source is 100, and that they are to use that value to estimate the brightness of other light sources. They are then presented with a series of light sources with different brightnesses, and are asked to write down their estimates for the values of these light sources. For example, if the participant believes that a given light source has half of the brightness of the standard, she would give it a value that is half of the value of the modulus, in this case 50. If the participant believes that a given light source is twice as bright as the standard, she would give it a value that is twice the modulus, in this case 200. The standard remains visible throughout the experiment.

Bard and colleagues (Bard et al. 1996) proposed a straightforward methodology for a type of magnitude estimation of acceptability. In ME of acceptability, participants are presented with a sentence (the standard) and a numeric value representing its acceptability (the modulus). They are then instructed to indicate the acceptability of all subsequent sentences using the acceptability of the standard:

Figure 5: An example of syntactic Magnitude Estimation

Standard:	Who thinks that my brother was kept tabs on by the FBI?
Acceptability:	100
Item:	What did Lisa meet the man that bought?
Acceptability:	

As in psychophysical ME, the standard in syntactic ME remains visible throughout the experiment. Magnitude estimation has increasingly gained currency in experimental linguistics

(Keller 2000, 2003, Featherston 2005a, 2005b, Sprouse 2009), although important questions remain about whether participants are able to make ratio-based judgments of acceptability (Sprouse and Cunningham *submitted*), and ultimately whether magnitude estimation is a more sensitive task than n-point scale tasks (Weskott and Fanselow 2009). Nevertheless, ME seems well suited to the present study because of its potential to capture finer-grained ratings using the infinite and unbounded positive number line.

5.1 Materials & Methods

Participants

176 self-reported monolingual native speakers of English, all University of California Irvine undergraduates, participated in this experiment for either course credit or $5 (152 Female). The experiment was administered during a single visit to the lab during which the participants completed the acceptability judgment task, the serial recall task, and the *n*-back task (in that order). Three participants were removed from analysis because they inverted the response scale in the acceptability task. All analyses below were run on the remaining 173 participants.

The acceptability rating task

Four island types were tested, each using a 2 × 2 manipulation of extraction and structural environment, as discussed in Section 2, yielding a total of 16 critical conditions: whether islands, complex NP islands, subject islands, and adjunct islands. Eight additional sentence types were included to add some variety to the materials, for a total of 24 sentence types. 16 lexicalizations of each sentence type were created, and distributed among 4 lists using a Latin Square procedure. This meant that each list consisted of 4 tokens per sentence type, for a total of 96 items. 2 orders for each of the 4 lists were created by pseudorandomizing the items such that related sentence types were never presented successively. This resulted in 8 different surveys. The standard was identical for all 8 surveys, and was in the middle range of acceptability: *Who said my brother was kept tabs on by the FBI?* The standard was assigned a modulus of 100. Example materials for Experiment 2 were structurally similar to those for Experiment 1 except for the subject island sub-design. The experimental materials are available on the first author's website.

(7) *Subject Island (for experiment 2 only)*

Who __ thinks the speech interrupted the primetime TV show?
What do you think __ interrupted the primetime TV show?
Who __ thinks the speech about global warming interrupted the primetime TV show?
What do you think the speech about __ interrupted the primetime TV show?

The acceptability rating task was presented as a paper survey. The experiment began with a practice phase during which participants estimated the lengths of 7 lines using another line as a standard set to a modulus of 100. This practice phase ensured that participants understood the concept of magnitude estimation. During the main phase of the experiment, 10 items were presented per page (except for the final page), with the standard appearing at the top of every page inside a textbox with black borders. The first 9 items of the survey were practice items (3

each of low, medium, and high acceptability). These practice items were not marked as such, i.e., the participants did not know they were practice items, and they did not vary between participants in order or lexicalization. Including the practice items, each survey was 105 items long. The task directions are available on the corresponding author's website. Participants were under no time constraints during their visit.

The serial recall task

The materials for the serial recall task consisted of the same 8 disyllabic words used in Experiment 1. The presentation of the serial recall task was identical to the presentation in Experiment 1, except for two minor changes. First, only 6 words (out of the pool of 8) were presented during each of the 10 trials. This change created a small amount of variation between trials, and hence made the task more interesting for participants. This change also brought the total number of words presented per trial within the range of recall scores observed in the first experiment in order to eliminate the possibility that the length of the trials impaired performance on the recall task. Second, participants were explicitly instructed to leave blanks on the response sheet when they could not remember the word that occurred in that position. This allowed us to use the standard scoring procedure described in Section 4 (Cowan 2000).

The n-back tasks

Each participant completed three *n*-back tasks: a 2-back, a 3-back, and a 4-back (in that order). The *n*-back tasks each consisted of a sequence of 30 letters drawn from a set of 8 possible letters. The sequence was different for each *n*, although the set of potential letters was identical. Ten of the letters in each sequence were potential *hits*, in the respect that they appeared n items after a previous presentation. Twenty of the letters in each sequence were potential *correct rejections*, in that they did not appear n items after a previous presentation. All participants saw the same sequences of items.

Participants performed the 2-back, 3-back, and 4-back in succession with a break between each task. The experiment was controlled by the DMDX presentation software (Forster and Forster 2003). The letter sequences were presented one at a time using a yellow font (size=48) on a blue screen. The letters were visible for 2s, with 1s of blank screen before the presentation of the next letter. Participants were instructed to press the green key (the J-key with a green cover) if they believed that the current item had appeared *n*-items previously. Participants were instructed to do nothing for all other cases. This is a standard feature of *n*-back tasks, and it aims to minimize distraction from the memory updating aspect of the task. Because there were 30 items per task, each task took approximately 90 seconds to complete.

Accuracy in this task was quantified using d′ sensitivity scores. The advantage of this measure is that it controls for response bias, so that it better reflects sensitivity to a given contrast. It does so by taking the z-score difference between *hits*, correct responses to *n*-back target-present trials, and *false alarms,* incorrect responses to target-absent trials (see MacMillan and Creelman 2004). The maximum possible d′ score was 3.6, indicating perfect discrimination. A d′ of 0 indicates chance-level discrimination, and a negative d′ score indicates worse than chance-level discrimination. Due to a computer problem early in the testing phase of experiment 2, 12 participants' *n*-back scores were corrupted. Therefore the sample size for the *n*-back analyses is 161 participants rather than 173 participants.

5.2 Results

As with Experiment 1, acceptability judgments from each participant were z-score transformed prior to analysis. The z-score transformation eliminates the influence of scale bias on the size of the DD scores, and therefore increases the likelihood of finding a significant relationship between working memory capacity and DD scores. DD scores were calculated using the formula presented in Section 3.

Table 6: Experiment 2, means and standard deviations of z-scored magnitude estimation scores for each condition (n=173)

	whether	complex NP	subject	adjunct
short, non-island	1.23 (0.74)	0.86 (0.76)	0.85 (0.77)	0.62 (0.80)
long, non-island	0.38 (0.72)	0.18 (0.82)	0.38 (0.83)	0.23 (0.79)
short, island	0.71 (0.67)	0.75 (0.71)	0.75 (0.79)	0.11 (0.81)
long, island	-0.73 (0.63)	-0.73 (0.57)	-0.97 (0.61)	-0.97 (0.72)

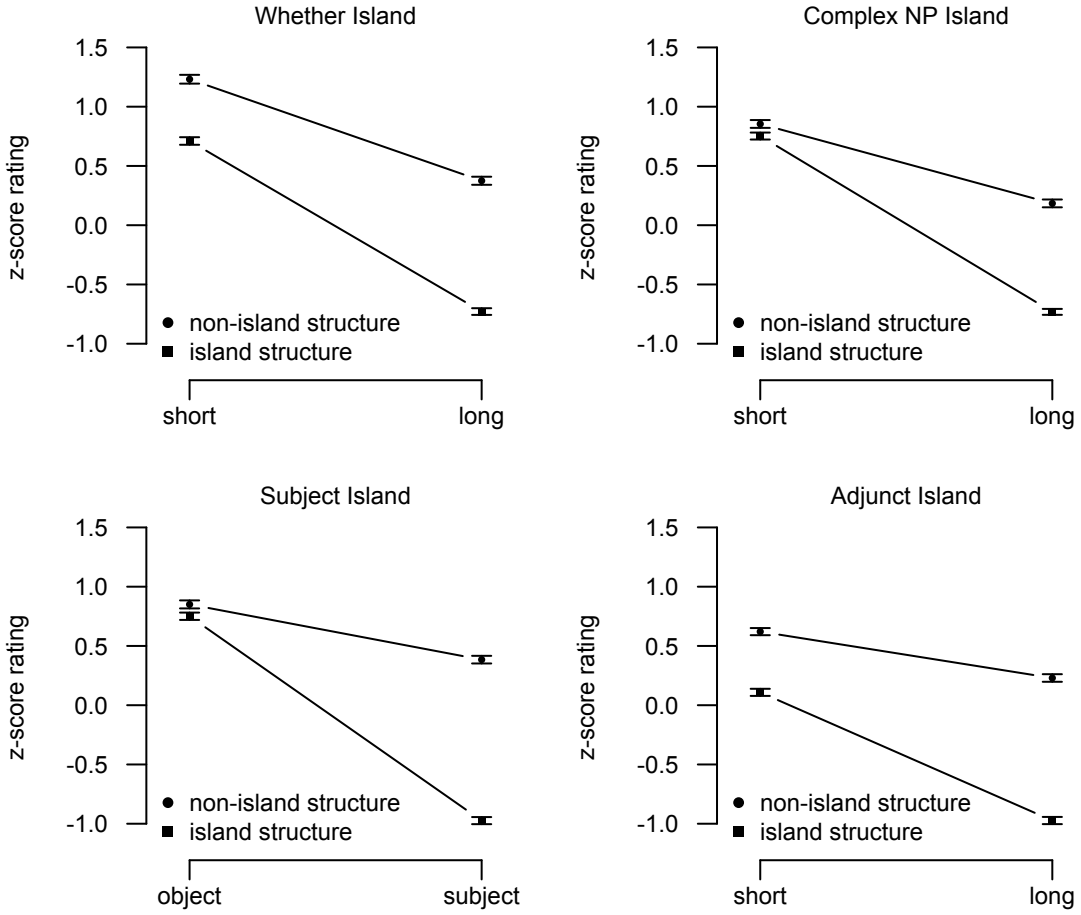
The basic island effects

Again, the first question we can ask is whether the basic island effects arise in our sample. Linear mixed effects models reveal a significant main effect of LENGTH for each island type. And unlike Experiment 1, this experiment showed a significant main effect of STRUCTURE for each island type. However, because the interactions are super-additive, it is possible that the long, island condition is driving these main effects. Therefore we also ran pairwise comparisons to isolate each of the potential processing costs. The length cost was isolated with a pairwise comparison of the short, non-island and long, non-island conditions. The structure cost was isolated with a pairwise comparison of the short, non-island and short, island conditions. As Table 7 indicates, the effect of LENGTH was significant for every island type as expected. However, the effect of STRUCTURE was not significant for complex NP and subject islands (even with our extremely large sample size of 173). This again raises the question of how island effects (the interaction) could be caused by the combination of two processing costs when the cost associated with island structures is only reliably present in the whether island, and is reliably absent in the complex NP island and the corrected subject island design.

Table 7: Experiment 2, two-way linear mixed effects models for each island type and pairwise comparisons for the effects of each structural manipulation (n=173)

	whether	complex NP	subject	adjunct
Main effect of LENGTH	.0001	.0001	.0001	.0001
Main effect of STRUCTURE	.0001	.0001	.0001	.0001
LENGTH x structure	.0001	.0001	.0001	.0001
Pairwise comparison: LENGTH	.0001	.0001	.0001	.0010
Pairwise comparison: STRUCTURE	.0001	.2260	.3514	.0001

Figure 6: Experiment 2, interaction plots for each island type (n=173)



Differences-in-differences as a function of serial recall

Serial recall scores ranged from 1.1 to 5.5, with a mean of 2.98 and a standard deviation of .80. As in Experiment 1 the acceptability ratings in Experiment 2 were z-score transformed prior to calculation of the DD scores. Linear regressions were performed for each island type using DD scores as the dependent variable, and serial recall scores as the independent variable, for both the complete set of DD scores and the set of DD scores above zero. The exclusion of DD scores below zero affected 36 scores for whether islands, 27 for CNPC islands, 17 for subject islands, and 31 for adjunct islands.

Figure 7: Experiment 2, differences-in-differences scores plotted as a function of serial recall scores (n=173). The solid line represents the line of best fit for all of the DD scores. The dashed line represents the line of best fit when DD scores below 0 are removed from the analysis (shaded grey). Trend lines were fitted using a least-squares procedure.

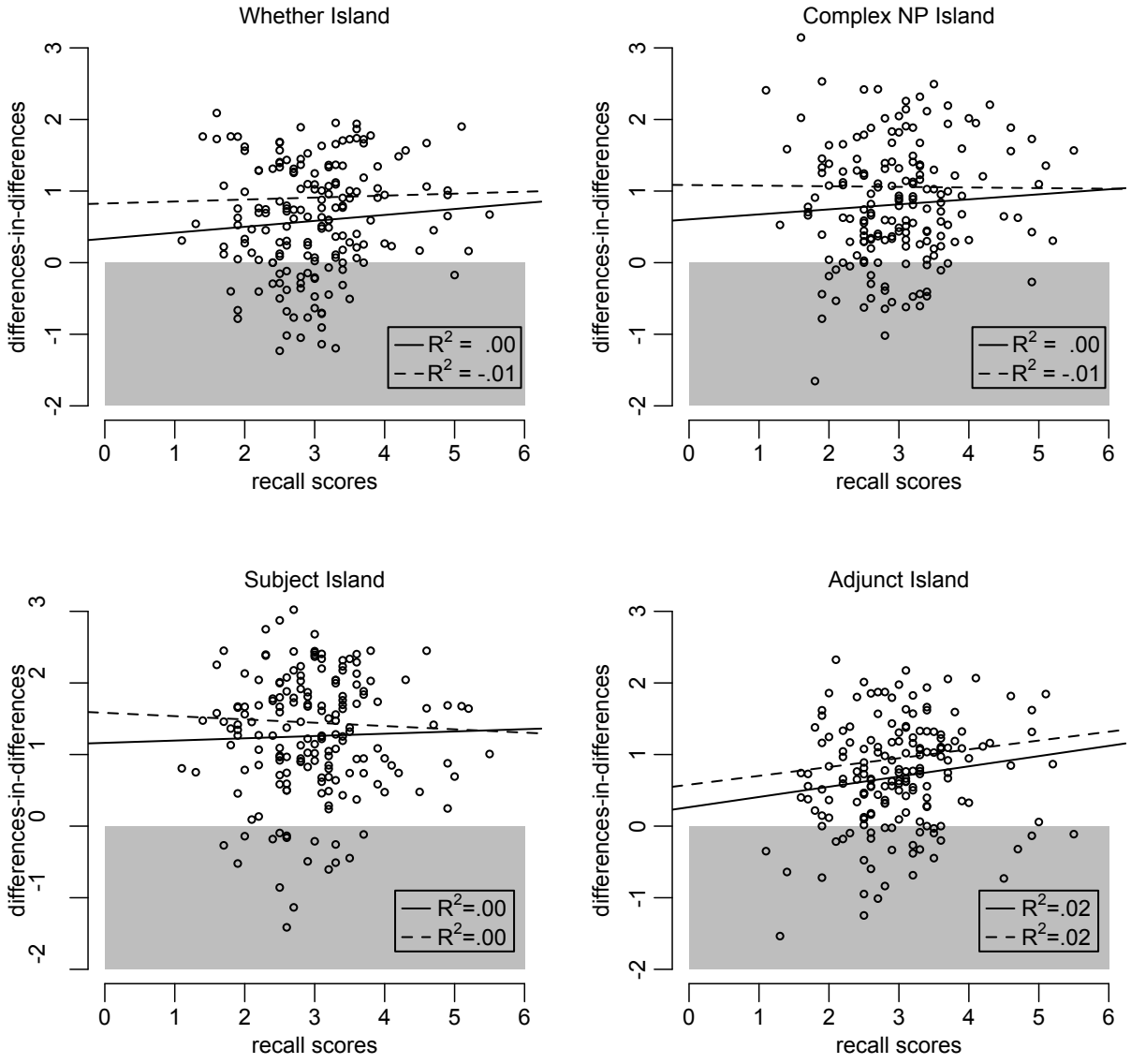


Table 8: Experiment 2, linear regression modeling differences-in-differences scores as a function of serial recall scores (n=173)

	island	intercept	slope of recall	t-statistic of recall	p-value of recall	adjusted R^2
All DDs	whether	0.34	0.08	1.05	.29	.00
	complex NP	0.60	0.07	0.88	.38	.00
	subject	1.16	0.03	0.39	.70	.00
	adjunct	0.26	0.14	2.02	.05	.02
DDs greater than zero	whether	0.83	0.03	0.48	.64	-.01
	complex NP	1.08	-0.01	-0.13	.90	-.01
	subject	1.58	-0.05	-0.71	.48	.00
	adjunct	0.58	0.12	2.02	.05	.02

As in Experiment 1, the models returned by the linear regression strongly suggest that there is no evidence of a meaningful relationship between DD scores and serial recall scores. R^2 values are extremely low – most are at or below zero, and the highest R^2 values were two conditions in which recall scores accounted for 2% of the variance in island effects. Negative R^2 values are possible with the adjusted R^2 measure used here (and in Experiment 1) because the adjustment for the number of degrees of freedom has the effect of lowering the standard R^2 value slightly. This adjustment is recommended because the standard R^2 value is a biased (inflated) statistic. Negative values suggest that the biased (inflated) R^2 statistic was at or near zero, and that the degrees-of-freedom correction brought the statistic below zero.

Differences-in-differences as a function of n-back

The descriptive results of the *n*-back tasks are reported in Table 9:

Table 9: Experiment 2, means and standard deviations for the *n*-back tasks (n=173)

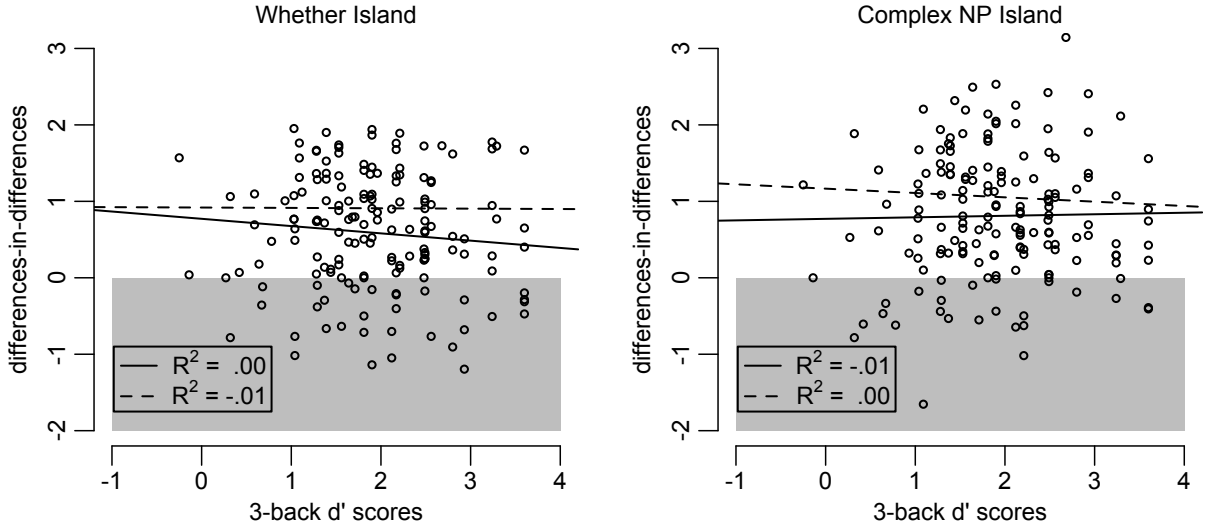
	minimum d′	maximum d′	mean	standard deviation
2-back	0.68	3.60	2.92	0.71
3-back	-0.25	3.60	1.92	0.81
4-back	-0.61	3.60	1.30	0.71

The analysis of the *n*-back scores is more complicated than the analysis of serial recall scores, since each participant contributed three d′ scores. In this subsection, we first present a linear regression using just the 3-back d′ scores as a predictor of DD scores. The benefit of this analysis is that the graph of the analysis is straightforward to plot, and the results are intuitive to interpret. The cost of this simplicity is that the analysis does not use all of the available *n*-back scores. Therefore we also present the results of a second analysis that includes all three scores *n*-back scores in a multiple linear regression. The benefit of this second model is that it is fully saturated, and therefore offers the strongest possibility of a good fit (large adjusted R^2). However, there are

two costs to this analysis. First, the model is 4-dimensional, so it cannot be plotted easily; therefore we only report the results in a table. Second, the three *n*-back scores that we wish to include in the model are necessarily correlated to some degree. Multiple linear regression requires that the factors be independent, so we must eliminate the correlation before constructing the model. A common method for eliminating correlation between factors (also known as collinearity) is to perform a Principal Component Analysis on the factors prior to the regression (Pearson 1901, Jolliffe 2002). Because we have three scores (2-back, 3-back, and 4-back), the result of the PCA is a set of three new factors, called components, that are completely uncorrelated with each other (i.e., the axes of the coordinate system of the new factors are orthogonal to each other), but that still capture the same variance as the original three factors. The independence of these three new factors (components) allows us to include them in the model without violating the assumptions of multiple linear regression.

We chose the 3-back as our illustration of a single predictor model because 48 participants performed perfectly on the 2-back task, suggesting that the task is too easy to be a representative memory score, and 7 participants performed below chance on the 4-back task, suggesting that the 4-back task may be too difficult to be a representative memory score. In contrast, only 7 participants performed perfectly and only 1 participant performed below chance on the 3-back task.

Figure 8: Experiment 2, differences-in-differences plotted as a function of 3-back scores (n=161). The solid line represents the line of best fit for all of the DD scores. The dashed line represents the line of best fit when DD scores below 0 are removed from the analysis (shaded grey). Trend lines were fitted using a least-squares procedure.



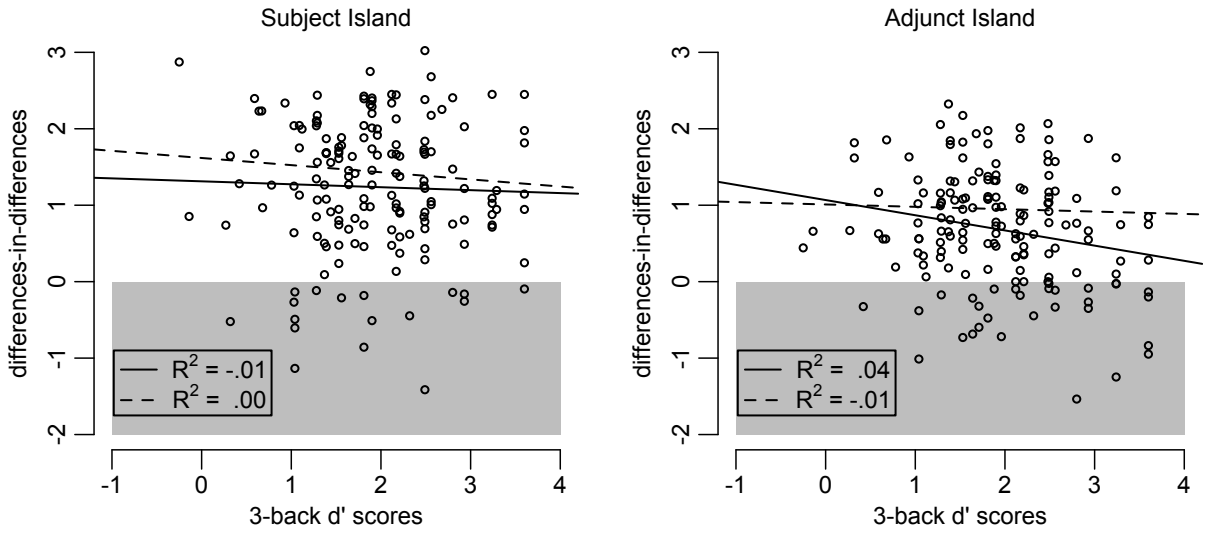


Table 10: Experiment 2, linear regression modeling differences-in-differences scores as a function of 3-back scores (n=161)

	island	intercept	slope of 3-back	t-statistic of 3-back	p-value of 3-back	adjusted R^2
All DDs	whether	0.77	-.10	-1.19	.24	.00
	complex NP	0.77	0.02	0.24	.81	-.01
	subject	1.31	-0.04	-0.45	.66	-.01
	adjunct	1.07	-0.20	-2.72	.01	.04
DDs greater than zero	whether	0.92	0.00	-0.07	.94	-.01
	complex NP	1.17	-0.06	-0.75	.45	.00
	subject	1.62	-0.09	-1.31	.19	.00
	adjunct	1.01	-0.03	-0.48	.63	-.01

The results of the linear regression on 3-back scores is similar to the results of the recall scores: 7 out of 8 of the models have adjusted R^2 values that are at or below 0, and the one model that is higher has an adjusted R^2 value of only .04. This suggests that there is no relationship between 3-back performance and island effects. For completeness we also ran simple linear regressions on the 2-back and 4-back scores respectively, with no change in the results.

For the model including all three *n*-back scores, we decided to include all three of the orthogonal components returned by the PCA in the model because each component accounted for a sizeable portion of the variance in the original data set: component 1 accounted for 45% of the original variance, component 2 accounted for 29% of the original variance, and component 3 accounted for 26% of the original variance. Including all three of the components in a saturated multiple linear regression model increases the likelihood that the model will capture a significant portion of the variance in DD scores, and thus increases the likelihood of finding support for the capacity-based theory. However, even with all three *n*-back scores included in the model, none of the models capture a meaningful amount of the variance (see the Appendix for the complete statistical details).

Table 11: Experiment 2, Adjusted R^2 values for a multiple linear regression analysis modeling differences-in-differences scores as a function of all three *n*-back scores after PCA (n=161)

	whether	complex NP	subject	adjunct
All DDs	.00	-.02	-.01	.03
DDs greater than zero	-.02	.01	-.01	-.02

As a final analysis, we created a model that includes both the serial recall scores and the *n*-back scores to see if including both types of working memory measures reveals a significant relationship with the strength of island effects. As with the three *n*-back scores, we first performed a PCA to eliminate any correlation between the four measures. The four components returned by the PCA explain the following proportions of variance respectively: 33%, 28%, 20%, 19%.

Table 12: Experiment 2, Adjusted R^2 for a multiple linear regression analysis modeling differences-in-differences scores as a function of all four memory scores after PCA (n=161)

	whether	complex NP	subject	adjunct
All DDs	.01	-.01	-.01	.05
DDs greater than zero	-.02	.00	-.01	.00

As Table 12 indicates, even when using a saturated model that includes both types of working memory scores, there is no evidence of a significant relationship between working memory and island effects. All of the adjusted R^2 values of the models are at or below 0, except for the adjunct island model that includes all of the DD scores, which has a slightly higher, but still very low, adjusted R^2 value of .05. It is also interesting to note that in the adjunct island model we see evidence that the model is affected by the negative DD scores, i.e., the scores from individuals who show a greater effect of extraction in non-island contexts than in island contexts. The slightly elevated adjusted R^2 of .05 found for the full data set reduces to zero when the negative DD scores are removed from the analysis.

Taken as a whole, the results of Experiment 2 provide no evidence of a relationship between working memory capacity and the strength of island effects. Furthermore, the parallel between the results of Experiment 1 and Experiment 2 suggests that the independence of working memory capacity and island judgments does not depend upon the choice of acceptability judgment task (7 point Likert scale vs. magnitude estimation) or the choice of working memory measure (serial recall vs. *n*-back).

6. A bootstrap-based simulation

The results of Experiments 1 and 2 suggest that there is no relationship between working memory capacity and the strength of island effects. However, because this conclusion is based upon a null effect, we conducted additional statistical analyses to further test this conclusion. In the previous sections we employed standard linear regression procedures to look for evidence of a relationship between working memory capacity and island effects. Although linear regression is generally considered to be a relatively sensitive tool, there was one step in our regression

analysis that discarded some of the information contained in our experimental results. In Experiment 2 each participant rated four tokens of each condition. However, to perform the regressions we combined those four ratings into a single average rating for each condition for each participant. In the process of averaging those four trials we eliminated some of the uncertainty in the rating process. Using a point-estimate in this manner is standard practice within (frequentist) behavioral statistics; however, it is logically possible that this averaging procedure could have obscured a relationship between working memory capacity and island effects in the data that we collected.

To investigate whether the averaging procedure obscured a potential significant relationship, we performed a bootstrap-based simulation using the data from Experiment 2. Broadly speaking, bootstrap simulations are used to derive an unbiased estimate of a population parameter from a given sample (under the assumption that the sample itself is representative of the population). For our purposes, we are interested in the adjusted R^2 of the relationship between working memory capacity and island effects. A bootstrap simulation provides us with more than just a single value given by the linear regressions in previous sections. Instead it provides a frequency distribution of possible values of the population parameter. Given a sufficient number of simulations (e.g., 10,000), this frequency distribution is a good approximation of a probability distribution of possible values of the population parameter. Furthermore, because the frequency distribution is a finite distribution, we can use it to identify the absolute highest R^2 possible, to see if there is any evidence whatsoever in our sample of a significant relationship between working memory capacity and island effects. In other words, assuming that our sample is representative of the population, the frequency distributions obtained by the bootstrap-based simulations can tell us: (i) the most likely R^2 value of the population, and (ii) the possible range of variation in the population.

For each island type (whether, CNPC, subject, and adjunct) and each working memory measure (serial recall and *n*-back), we performed the following steps:

1. For each condition for each participant, we took a sample of size 4 (with replacement) from the 4 judgments given by the participant. This means that at one limit this new sample could potentially have the same 4 judgments as the participant actually gave, or at the other limit it could have one of the judgments repeated 4 times. In other words, this sampling procedure returns a new set of 4 judgments within the bounds of uncertainty delineated by the original 4 judgments.

2. We took the average of the 4 scores in the new sample and saved it as one possible score for that condition and that participant.

3. We repeated Steps 1-2 10,000 times for each condition (n=16) and each participant (n=173) to create 10,000 possible scores for each condition for each participant.

4. Next we calculated a DD score for each island type for each participant using the 10,000 scores we obtained in Step 3. This resulted in 10,000 DD scores for each participant.

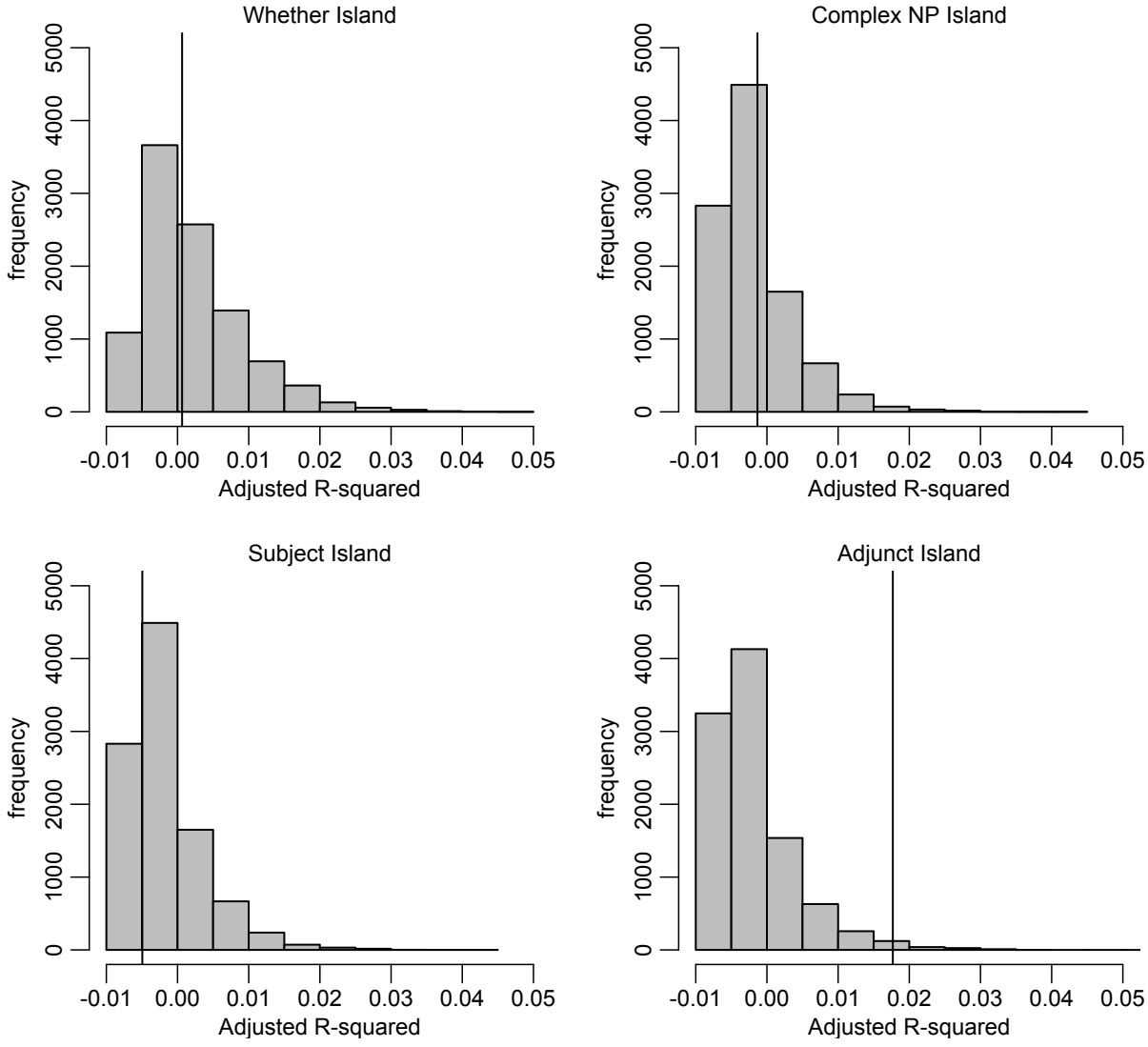
5. We then chose one of the DD scores for each participant, and input those DD scores into a linear regression. Again, this is just like our original analysis, as it is a linear regression based on 173 pairs of DD scores and WM scores (161 DD scores in the case of *n*-back).

The difference is that the DD scores were chosen from the simulated DD scores derived from the previous steps.

6. We repeated Step 5 for each of 10,000 simulated DD scores. This resulted in 10,000 total linear regressions based on the simulated data.

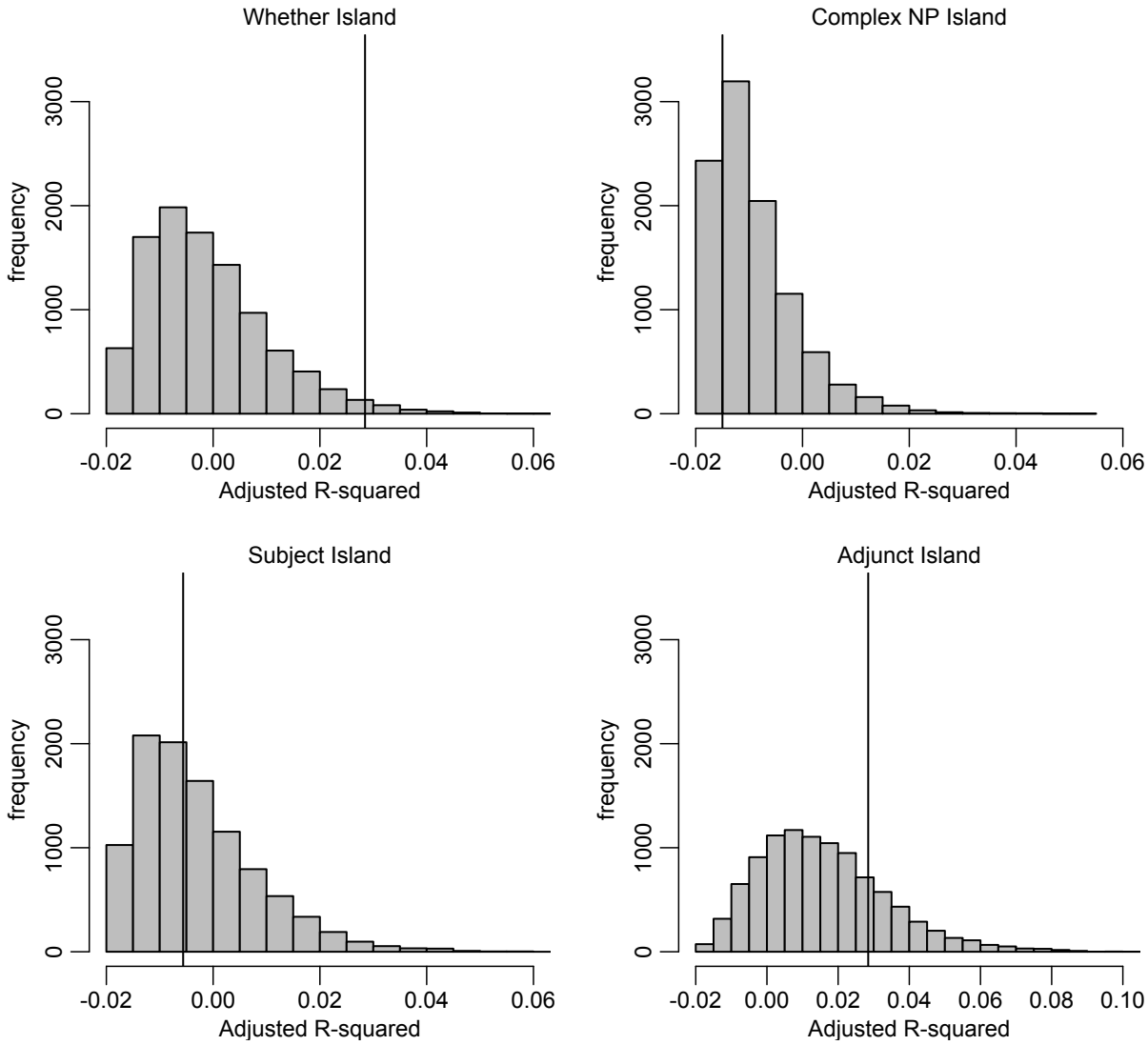
Finally, we plotted the distribution of the adjusted R^2 values for those 10,000 linear regressions to see if there is ever any evidence of a significant relationship between DD scores and working memory capacity.

Figure 9: Simulation study of experiment 2, histograms of adjusted R^2 values for linear regressions modeling differences-in-differences scores as a function of serial recall scores (n=10,000). Distribution of adjusted R^2 across the 10,000 simulations. The solid line represents the R^2 value of the actual sample (analyzed in section 4).



The frequency distribution of R^2 values for serial recall scores corroborates the results from Section 5: the most probable population value is at or below 0, and across all 10,000 simulated data sets the R^2 value is consistently below .05, providing further evidence that there is no significant relationship between serial recall performance and the strength of island effects.

Figure 10: Simulation study of experiment 2, histograms of adjusted R^2 values for linear regressions modeling differences-in-differences scores as a function of all three n-back scores (n=10,000). Distribution of adjusted R^2 across the 10,000 simulations. The solid line represents the R^2 value of the actual sample (analyzed in section 4).



The frequency distribution of adjusted R^2 values for *n*-back scores also corroborates the results from Section 5: the most probable population value is at or below 0 for all of the island types except the adjunct island, which exhibits a slightly wider distribution that centers around an R^2

value of .01. The range of variation is below .06 for all island types except for the adjunct island, which has a slightly wider distribution that reaches an R^2 value of .10 in the extreme. These simulations provide further evidence that there is no significant relationship between *n*-back scores and island effects. Taken together, these results confirm that it is unlikely that the null effect found in Section 5 was due to the averaging procedure required by the linear regression analysis, and to the extent that our sample is representative of the larger population, it is unlikely that there is a significant relationship between working memory capacity and island effects.

7. Consequences for the capacity-based theory

In this article, we have argued that the key difference between the capacity-based reductionist theory and grammatical theories lies in the how they interpret the statistical interaction in acceptability scores that arises when dependency length and structure are independently manipulated: the capacity-based theory analyzes the statistical interaction as the consequence of a psychological interaction of two (sets of) processes due to limited capacity, whereas grammatical theories analyze the statistical interaction as an consequence of a constraint that impacts judgments on only one of the four conditions in the design. Therefore the capacity-based theory predicts that the interaction should correlate with capacity measures, whereas grammatical theories predict that the interaction should not correlate with capacity measures.

In Sections 4 and 5 we presented two large-scale studies, with 142 and 173 participants each, that were designed to test for a relationship between the strength of the interaction and processing resource capacity. We used two different response scales for the acceptability judgment task (7-point and magnitude estimation), and two different types of working memory measures (serial recall and *n*-back), but found no evidence of a relationship between the statistical interaction and resource capacity. Furthermore, we conducted a bootstrap-based simulation on our data to ensure that the relationship was not obscured by the averaging procedure used in our original analyses, but we still found no evidence of a relationship between the strength of the interaction and resource capacity. In fact, for complex NP and subject islands we did not even find evidence of the processing cost of the island structure, contradicting one of the premises of the capacity-based theory. These results are consistent with grammatical theories of island effects, which predict no relation between resource capacity and the strength of island effects. The results are also compatible with 'grounded' theories, which posit a role for resource capacity constraints in the history or evolution of island constraints, but assume that the constraints are explicitly represented as formal constraints in the minds of contemporary speakers. However, these are results are incompatible with a capacity-based reductionist theory.

One potential objection to this interpretation of the results could be that we only tested two capacity measures, so it is logically possible that a different capacity measure could be found that does indeed correlate with the strength of island effects. While this objection is logically sound, it is undermined by the fact that most working memory measures are highly correlated (Conway et al. 2005). This objection would hence require a working memory measure that is simultaneously highly correlated with island effects, but not at all correlated with serial recall or *n*-back, otherwise the lack of effect found in the current experiments would go unexplained. However, given that recent work suggests that the serial recall and *n*-back tasks are uncorrelated (Kane et al. 2007), the likelihood of finding a new measure that correlates with neither is very small indeed. Of course, it is also possible to counter the present findings with the argument that the processing resources relevant to island effects are simply not measureable. Considerable

caution should be observed in pursuing that explanation. Such an approach could certainly be used to discount the results presented here, but it would also have the unintended consequence of discounting the primary attraction of the capacity-based theory. The attraction of reductionist theories is that they eliminate the need for cognitive constructs that are only postulated to explain island effects. If the "processing resources" necessary to explain island effects are not independently motivated, then there is no advantage to postulating them over a grammatical constraint.

Another potential objection to this interpretation could be that we have focused on the wrong population: university students may not show the necessary variation in memory capacity necessary to reveal variation in island effects. Specifically, it may be the case that these samples simply did not sample a sufficient number of participants from the high end of the working memory scale. Crucially, the distribution of serial recall scores in Experiment 2 (which used the standard scoring procedure) are consistent with the distribution of scores reported in the meta-analysis in Cowan 2000. Therefore this objection must reduce to a concern about sampling bias in university-based working-memory experiments generally. One possible response to this is that professional linguists are rather uniform in their judgments of the unacceptability of island effects in English. Under the plausible assumption that professional linguists as a population have higher sentence processing capacity than the general population, this would suggest that there may be no English speakers with enough capacity to successfully parse island effects. It is not clear if this is a claim that proponents of the capacity-based theory would be comfortable endorsing.

One final objection to the results of this study could be that some theories of filler-gap dependency completion reject the claim that maintenance of the wh-word is required (Fodor 1978, McElree et al. 2003). This could suggest that the best candidate processes for a capacity-based theory are not those associated with maintaining a *wh*-word in working memory (as suggested by Kluender and Kutas 1993), but rather those associated with retrieving that *wh*-word from working memory at the gap location. It is logically possible that maintenance capacity and retrieval capacity could show independent variation within the population, and that the capacity measures used in this study tracked the former rather than the latter capacity. Again, the facts of memory capacity measures make this highly unlikely, as it is known that capacity measures, as an experimental construct, likely tap into individual differences related to both maintenance and retrieval (such as, interference robustness (Unsworth & Engle, 2007)). Therefore the capacity-based theory may shift its emphasis on component memory processes, but the tasks necessary to measure capacity for those processes are unlikely to differ sufficiently from the ones employed in the present study.

In sum, we believe that the results of the large-scale experiments presented in this article provide strong support for grammatical theories of island effects, as we can find no evidence of a relationship between processing resource capacity and island effects. And while we can envision several potential objections to this interpretation, the known facts of working memory tasks suggest that the likelihood of finding such a relationship with different tasks or a different sample is extremely small. These results suggest that the most profitable avenues available for furthering our understanding of island effects are grounded theories, which maintain many of the attractive aspects of reductionist theories, and grammatical theories, which are actively being research by many linguists.

References

Baayen, R. H. 2007. Analyzing linguistic data: a practical introduction to statistics using R. Cambridge University Press.

Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59.390-412.

Bard, E. G., D. Robertson, and A. Sorace. 1996. Magnitude estimation of linguistic acceptability. Language 72.32-68.

Berwick, R. C. and A. S. Weinberg. 1984. The grammatical basis of linguistic performance. Cambridge, MA: The MIT Press.

Caplan, D. and G. S. Waters. 1999. Verbal working memory and sentence comprehension. Behavioral and Brain Sciences. 22(1).77-94.

Chomsky, N. 1973. Conditions on transformations. A Festschrift for Morris Halle, ed. by S. Andersonand and P. Kiparsky. New York: Holt, Rinehart, and Winston.

Chomsky, N. 1986. Barriers. Cambridge, MA: The MIT Press.

Cowan, N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences 24.87-114.

Conway, A. R. A., M. J. Kane, M. F. Bunting, D. Z. Hambrick, W. Oliver, and R. W. Engle. 2005. Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin & Review 12(5).769-786.

Daneman M. and P. A. Carpenter. 1980. Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior 19.450-466.

Deane, P. 1991. Limits to attention: a cognitive theory of island phenomena. Cognitive Linguistics 2.1-63.

Erteschik-Shir, N. 1973. On the nature of island constriaints. Cambridge, MA: MIT dissertation.

Featherston, S. 2005a Magnitude estimation and what it can do for your syntax: some wh-constraints in German. Lingua 115.1525-1550.

Featherston, S. 2005b. Universals and grammaticality: wh-constraints in German and English. Linguistics 43.667-711.

Fedorenko, E., Gibson, E. & Rohde, D. 2006. The Nature of Working Memory Capacity in Sentence Comprehension: Evidence against Domain Specific Resources. Journal of Memory and Language, 54(4).541-553.

Fedorenko, E., Gibson, E. & Rohde, D. 2007. The nature of working memory in linguistic, arithmetic and spatial integration processes. Journal of Memory and Language 56(2).246-269.

Fodor, J. D. 1978. Parsing strategies and constraints on transformations. Linguistic Inquiry 9.427-473.

Fodor, J. D. 1983. Phrase structure parsing and the island constraints. Linguistics and Philospohy 6. 163-223.

Forster, K. I. and J. C. Forster 2003. DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods 35.116-124.

Goldberg, A. 2007. Constructions at work. Oxford University Press.

Hawkins, J. A. 1999. Processing complexity and filler-gap dependencies across grammars. Language 75.244-285.

Hofmeister, P. and I. A. Sag. 2010. Cognitive constraints and island effects. Language 86.

Jaeggi, S. M., M. Buschkuehl, J. Jonides, and W. Perrig. 2008. Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences 105(19).6829–6833.

Jolliffe, I. T. 2002. Principal Component Analysis. New York: Springer-Verlag.

Just, M. A., and P. A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. Psychological Review 98.122–149.

Kane, M. J., & R. W. Engle. 2002. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual differences perspective. Psychonomic Bulletin & Review 9. 637-671.

Kane, M. J., A. R. A. Conway, T. K. Miura, and J. H. Colfesh. 2007. Working memory, attention control, and the *n*-back task: A question of construct validity. Journal of Experimental Psychology: Learning, Memory, and Cognition 33(3).615–622.

Keller, F. 2000. Gradience in grammar: experimental and computational aspects of degrees of grammaticality. Edinburgh: University of Edinburgh dissertation.

Keller, F. 2003. A Psychophysical Law for Linguistic Judgments. Proceedings of the 25th Annual Conference of the Cognitive Science Society, ed. by R. Alterman and D. Kirsh, 652-657. Boston.

King, J., and M. A. Just. 1991. Individual differences in syntactic processing: the role of working memory. Journal of Memory and Language 30.580-602.

Kirchner, W. K. 1958. Age differences in short-term retention of rapidly changing information. Journal of Experimental Psychology 55(4).352-358.

Kluender, R., and M. Kutas. 1993. Subjacency as a processing phenomenon. Language and Cognitive Processes 8.573-633.

Kluender, R. 1998. On the distinction between strong and weak islands: a processing perspective. Syntax and Semantics 29:The Limits of Syntax.241-279.

Kluender, R. 2004. Are subject islands subject to a processing account? Proceedings of the West Coast Conference on Formal Linguistics 23, ed. by V. Chand, A. Kelleher, A. J. Rodriguez, and B. Schmeiser, 101–125. Somerville, MA: Cascadilla Press.

Huang, C-T. J. 1982. Logical relations in Chinese and the theory of grammar. Cambridge, MA: MIT dissertation.

MacDonald, M. C., M. A. Just, and P. A. Carpenter 1992. Working memory constraints on the processing of syntactic ambiguity. Cognitive Psychology 24.56–98.

Macmillan, N. A. and C. D. Creelman. 2004. Detection Theory: A User's Guide. Mahwah, N.J.: Lawrence Erlbaum Associates.

S. E. Maxwell & H. D. Delaney 2003. Designing Experiments and Analyzing Data: A model comparison perspective. Mahwah, N.J.: Lawrence Erlbaum Associates.

McElree, B., S. Foraker, and L. Dyer. 2003. Memory structures that subserve sentence comprehension. Journal of Memory and Language 48.67-91.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 6(2).559-572.

Roberts, R. and E. Gibson. 2002. Individual differences in working memory. Journal of Psycholinguistic Research 31(6).573-598.

R Development Core Team. 2009. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.

Ross, J. 1967. Constraints on variables in syntax. Cambridge, MA: MIT dissertation.

Sprouse, J. 2007. A program for experimental syntax. College Park, MD: University of Maryland dissertation.

Sprouse, J. 2009. Revisiting Satiation. Linguistic Inquiry. 40(2).329-341.

Sprouse, J., S. Fukuda, H. Ono, and R. Kluender. in press. Grammatical operations, parsing processes, and the nature of wh-dependencies in English and Japanese. Syntax.

Sprouse, J. and H. Cunningham. (submitted). Evaluating the assumptions of magnitude estimation of linguistic acceptability. Language.

Stevens, S. S. 1956. The direct estimation of sensory magnitudes: loudness. The American journal of psychology 69.1-25.

Szabolcsi, A. and M. den Dikken. 2006. Strong and weak islands. In Martin Everaert and Henk van Riemsdijk, eds, The Blackwell Companion to Syntax. Mouton de Gruyter.

Szabolcsi A. and F. Zwarts. 1993. Weak islands and an algebraic semantics of scope taking. Natural Language Semantics 1(3).235-284.

Truswell, R. 2007. Extraction from adjuncts and the structure of events. Lingua 117.1355–1377.

Ueno, M. and R. Kluender. 2009. On the processing of Japanese wh-Questions: An ERP study. Brain Research 1290.63-90.

Unsworth, N. & Engle, Randall W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. Psychological Review 114.104-132.

Vos, S. H., Gunter, T. C., Schriefers, H. & Friederici, A. D. (2001). Syntactic parsing and working memory: The effects of syntactic complexity, reading span, and concurrent load. Language and Cognitive Processes 16(1).65-103.

Weskott, T., and G. Fanselow. 2009. Variance and informativity in different measures of linguistic acceptability. Proceedings of the 27th West Coast Conference on Formal Linguistics, ed. by N. Abner and J. Bishop, 431-439. Somerville, MA: Cascadilla Press.

Appendix of statistical results

Table A1: Experiment 2, Whether islands, multiple linear regression including all three n-back scores after PCA

		coefficient	t-statistic	p-value	adjusted R^2	F statistic
All DDs	Intercept	0.59	9.23	.001	.00	0.87
	Comp1	0.12	1.52	.13		
	Comp2	-0.03	-0.29	.76		
	Comp3	0.04	0.45	.65		
DDs greater than zero	Intercept	0.91	18.30	.001	-.02	0.16
	Comp1	0.02	0.26	.80		
	Comp2	-0.02	-0.27	.79		
	Comp3	0.04	0.57	.57		

Table A2: Experiment 2, Complex NP islands, multiple linear regression including all three n-back scores after PCA

		coefficient	t-statistic	p-value	adjusted R^2	F statistic
All DDs	Intercept	0.81	12.25	.001	-.02	0.21
	Comp1	0.00	0.05	.96		
	Comp2	-0.06	-0.57	.57		
	Comp3	0.06	0.56	.58		
DDs greater than zero	Intercept	1.06	18.51	.001	.01	1.36
	Comp1	0.05	0.67	.51		
	Comp2	0.05	0.65	.52		
	Comp3	0.16	1.85	.07		

Table A3: Experiment 2, Subject islands, multiple linear regression including all three n-back scores after PCA

		coefficient	t-statistic	p-value	adjusted R^2	F statistic
All DDs	Intercept	1.24	18.07	.001	-.01	0.70
	Comp1	-0.02	-0.27	.76		
	Comp2	0.14	1.39	.17		
	Comp3	-0.04	-0.33	.74		
DDs greater than zero	Intercept	1.44	25.61	.001	-.01	0.66
	Comp1	0.06	0.88	.38		
	Comp2	0.09	1.10	.28		
	Comp3	-0.01	-0.14	.89		

Table A4: Experiment 2, Adjunct islands, multiple linear regression including all three n-back scores after PCA

		coefficient	t-statistic	p-value	adjusted R^2	F statistic
All DDs	Intercept	0.68	11.78	.001	.03	2.56
	Comp1	0.14	2.06	.04		
	Comp2	0.16	1.86	.07		
	Comp3	0.00	0.09	.97		
DDs greater than zero	Intercept	0.95	19.37	.001	-.02	0.27
	Comp1	0.04	0.64	.52		
	Comp2	-0.02	-0.26	.80		
	Comp3	-0.04	-0.54	.59		

Table A5: Experiment 2, Whether islands, multiple linear regression including all four memory scores after PCA: serial recall, 2-back, 3-back, and 4-back

		coefficient	t-statistic	p-value	adjusted R^2	F statistic
All DDs	Intercept	0.59	9.27	.001	0.01	1.27
	Comp1	-0.06	-0.87	.38		
	Comp2	0.17	1.99	.05		
	Comp3	-0.04	-0.38	.70		
	Comp4	-0.05	-0.45	.65		
DDs greater than zero	Intercept	0.91	18.26	.001	-0.02	0.44
	Comp1	0.01	0.14	.89		
	Comp2	0.07	1.16	.25		
	Comp3	-0.01	-0.17	.87		
	Comp4	-0.05	-0.61	.54		

Table A6: Experiment 2, Complex NP islands, multiple linear regression including all four memory scores after PCA: serial recall, 2-back, 3-back, and 4-back

		coefficient	t-statistic	p-value	adjusted R^2	F statistic
All DDs	Intercept	0.81	12.29	.001	-.01	0.61
	Comp1	0.04	0.47	.64		
	Comp2	0.12	1.39	.17		
	Comp3	0.00	0.02	.99		
	Comp4	-0.06	-0.56	.58		
DDs greater than zero	Intercept	1.06	18.41	.001	.00	1.01
	Comp1	-0.05	-0.66	.51		
	Comp2	0.00	0.01	.99		
	Comp3	-0.06	-0.63	.53		
	Comp4	-0.16	-1.83	.07		

Table A7: Experiment 2, Subject islands, multiple linear regression including all four memory scores after PCA: serial recall, 2-back, 3-back, and 4-back

		coefficient	t-statistic	p-value	adjusted R^2	F statistic
All DDs	Intercept	1.24	18.04	.001	-0.01	0.66
	Comp1	0.03	0.37	.71		
	Comp2	0.00	0.04	.97		
	Comp3	-0.16	-1.55	.12		
	Comp4	0.03	0.32	.75		
DDs greater than zero	Intercept	1.44	25.53	.001	-0.01	0.55
	Comp1	-0.07	-1.09	.28		
	Comp2	-0.03	-0.50	.62		
	Comp3	-0.07	-0.87	.39		
	Comp4	0.01	0.13	.90		

Table A8: Experiment 2, Adjunct islands, multiple linear regression including all four memory scores after PCA: serial recall, 2-back, 3-back, and 4-back

		coefficient	t-statistic	p-value	adjusted R^2	F statistic
All DDs	Intercept	0.69	11.90	.001	0.05	3.07
	Comp1	-0.10	-1.42	.16		
	Comp2	0.14	1.85	.07		
	Comp3	-0.23	-2.61	.01		
	Comp4	0.00	0.03	.97		
DDs greater than zero	Intercept	0.95	19.45	.001	0.00	0.93
	Comp1	0.00	-0.06	.95		
	Comp2	0.12	1.82	.07		
	Comp3	-0.04	-0.49	.62		
	Comp4	0.04	0.48	.63		