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

Jon Sprouse
Department of Cognitive Sciences
University of California, Irvine
jsprouse@uci.edu

Matt Wagers
Department of Linguistics
University of California, Santa Cruz
mwagers@ucsc.edu

Colin Phillips
Department of Linguistics
University of Maryland, College Park
colin@umd.edu

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 analyze island effects as emergent consequences of non-grammatical constraints on the sentence parser, such as limited processing resources. In this paper we present two studies designed to test a fundamental prediction of one of the most prominent reductionist theories: that the strength of island effects should vary across speakers as a function of individual differences in processing resources. We tested over 300 native speakers of English on four different island effect types (Whether, Complex NP, Subject, and Adjunct Islands) 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 statistical analysis techniques, including resampling simulations. These results suggest that island effects are more likely to be due to grammatical constraints or grounded grammaticized constraints than to limited processing resources.

KEYWORDS: syntax, island constraints, acceptability judgments, working memory, language processing resources, individual differences

*This work was supported in part by National Science Foundation grants to JS (BCS-0843896) and CP (BCS-0848554). We would like to thank the following people for helpful comments and suggestions on earlier versions of this work: the audiences at GLOW 32, WCCFL 2010, UCLA, UCSD, the University of Chicago, and the École Normale Supérieure, Jim McCloskey, and two anonymous *Language* reviewers. All errors remain our own.

1. INTRODUCTION. Many of the world's languages exhibit constructions that contain a long-distance dependency between two elements in the sentence. For example, the English wh-questions in (1) illustrate such a dependency between the wh-phrase at the beginning of the sentence and the argument position of an embedded verb, indicated by a gap. Long-distance dependencies possess 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. WHETHER ISLAND
*What do you wonder [whether John bought __]?
b. COMPLEX NP ISLAND
*What did you make [the claim that John bought __]?
c. SUBJECT ISLAND
*What do you think [the speech about __] interrupted the TV show?
d. ADJUNCT ISLAND
*What do you worry [if John buys __]?
e. RELATIVE CLAUSE ISLAND
*What did you meet [the scientist who invented __]?
f. SENTENTIAL SUBJECT ISLAND
*What did [that John wrote __] offend the editor?
g. COORDINATE STRUCTURE ISLAND
*What did John buy [a shirt and __]?
h. LEFT-BRANCH ISLAND
*Which did John borrow [__ book]?

Many researchers have taken these facts to suggest that human grammars contain complex structural constraints on the rules that create long-distance dependencies. These constraints are referred to as island constraints, after Ross (1967). This grammatical approach to explaining the patterns in (1) and (2) has had 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 long-distance dependencies that are nonetheless constrained by specific structural configurations. In this way, island effects provide a classic motivation for theories that assume domain-specific constraints on language acquisition (i.e. Universal Grammar).

Given the far reaching consequences of the grammatical approach to island effects, it is perhaps unsurprising that there is a second class of theories – which we will call reductionist theories – that explicitly reject the conclusion that island constraints are part of the contents of grammars. Reductionist theories argue that some or all island effects can be reduced to independently motivated constraints on the functioning of the human sentence processor (Givón

1979, Deane 1991, Pritchett 1991, Kluender and Kutas 1993, Kluender 1998, Kluender 2004, Hofmeister and Sag 2010). According to reductionist theories, speakers do not explicitly represent (all) island constraints on the formation of long-distance dependencies. Instead, the perception of unacceptability arises as a by-product of the processing requirements of the sentence. An important theoretical attraction of such an approach is that it raises the promise 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. Therefore our goal in this article is to attempt to tease apart the grammatical and reductionist approaches to island effects by testing the role processing resource capacity plays in the perception of unacceptability of island-crossing long-distance dependencies. We do so by examining the relation between individual differences in working memory and how severe individuals perceive island violations to be.

Before delving into the details of our experiments, we should make several of our starting assumptions clear. First, we focus here on a single type of long-distance dependency (wh-dependencies in English) and four island types: Whether islands (2a), Complex NP islands (2b), Subject islands (2c), and Adjunct islands (2d). However, it should be noted that island effects have been observed with many different structures, such as 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), and many different types of long-distance dependencies, such as relativization (3a), topicalization (3b), and adjective-though constructions (3c), to name but a few.

(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.

Second, for the purposes of this study we collapse all grammatical approaches to islands into a single class, because they all hold that island constraints are independent of processing resources. However, it should be noted that there is substantial variability within the class of grammatical approaches. There are syntactic approaches to island effects, which posit syntactic constraints known as island constraints on the formation of wh-dependencies, such as Ross's classic Complex NP Constraint (Ross 1967), Chomsky's Subjacency Condition (Chomsky 1973,

[1] The purported learnability benefit of reductionist theories is not as straightforward as is often assumed. First, Pearl and Sprouse (2011) have recently argued that syntactic island constraints can indeed be learned from child-directed speech using a statistical learner with no innate, domain-specific learning biases. This suggests that there may be little to no consequences to the learnability problem by assuming a grammatical approach to island effects. Second, shifting the burden from the theory of grammar to a theory of processing costs may in fact add a learnability problem rather than eliminate one, as such an approach 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.

1986), and Huang's Condition on Extraction Domains (Huang 1982). There are also semantic approaches to island effects, for example, the algebraic semantics approach to weak islands by Szabolcsi and Zwarts (1993), the event structure account of Adjunct islands by Truswell (2007), and the presupposition failure account of negative islands, factives islands and others by Abrusán (2011). Some other grammatical approaches are more pragmatic in nature (Kuno 1976, Erteschik-Shir 1979, Kuno 1987, Kuno & Takami 1993, Goldberg 2007).

Third, we focus solely on the acceptability of classic island effects (and their relationship with working memory capacity), thus ignoring several other facets of island effects that may be relevant to adjudicating between grammatical approaches and reductionist approaches, such as the existence of island effects without displacement of the wh-word (i.e. wh-in-situ: Huang 1982, Lasnik and Saito 1984), the amelioration of island effects when a second gap is added to the sentence (i.e. parasitic gaps: Engdahl 1983, Phillips 2006), and the constrained cross-linguistic variation in island effects (e.g. Rizzi 1982, Torrego 1984). A complete theory of island effects, be it grammatical or reductionist, must also account for these phenomena.

Fourth, although there is also considerable variability within reductionist approaches to island effects, we focus exclusively the resource-limitation theory first proposed by Kluender and Kutas 1993 (and expanded in Kluender 1998, 2004, and Hofmeister and Sag 2010). We choose this particular reductionist theory for two reasons: (i) it is the most prominent reductionist approach in the literature, and (ii) its mechanisms are the most well-defined of all of the reductionist theories.

Finally, it should be noted that there is a third approach to island effects that in many ways represents a middle ground between grammatical and reductionist approaches: we call these grounded approaches. Grounded approaches share with grammatical approaches the assumption that the immediate cause of island effects in a speaker's mind is a formal grammatical constraint, and they share with reductionist approaches the assumption that island effects are ultimately a consequence of independently motivated properties of the human sentence parser. In this way, grounded approaches argue that island effects could have arisen historically because of parsing efficiency considerations, but that this efficiency consideration was ultimately grammaticized as a set of explicitly represented constraints (e.g. 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 identical (synchronic) predictions as grammatical theories, at least with respect to the role of processing resource capacity in the perception of unacceptability of island effects.

In what follows we first discuss the crucial distinction between resource-limitation theories and grammatical theories. Second, we present the results of two studies designed to test those predictions directly, and discuss their consequences for resource-limitation reductionist theories. The rest of this article is organized as follows. In Section 2 we discuss the resource-limitation theory in detail in an attempt to identify divergent predictions of the resource-limitation theory and grammatical theories. We argue that divergent predictions arise only when island effects are operationalized 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 studies 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. 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 in those analyses. Again, the results suggest no evidence of a relationship between resource capacity and island effects. We interpret these findings as inconsistent with the resource-limitation 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 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 treats the effect of capacity constraints in evolutionary terms, rather than synchronically; or the grammatical approach.

2. REDUCTIONISM AND THE RESOURCE-LIMITATION THEORY OF ISLAND EFFECTS.

2.1. THE SIMPLEST REDUCTIONIST THEORY OF ISLAND EFFECTS. The central claim of reductionist theories 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). In order to construct a basic reductionist theory, we can simply combine this claim with a few plausible assumptions about the human sentence processing system (although as we will see shortly, this simple implementation must be refined to capture the empirical facts of island effects).

(4) Assumptions of the simplest reductionist theory of island effects

Component 1: There is a processing cost associated with the operations necessary to build long-distance wh-dependencies

Component 2: There is a processing cost associated with the operations necessary to build the island structures

Linking hypothesis: Processing costs are reflected in acceptability judgments such that higher costs lead to lower acceptability

We have thus two potentially independently motivated processing costs -- the cost of long-distance dependencies and the cost of the island structure -- and we have a linking hypothesis between processing costs and acceptability judgments. Further we assume that each individual processing cost is small enough that sentences exacting only one of the costs are still considered acceptable. However, when both are combined in a single sentence, the sum of the two costs is large enough to cause the sentence to cross some threshold of unacceptability that separates acceptable sentences from unacceptable sentences. Finally, we must assume that this happens consistently across the different types of island structures in such a way that the combined cost for any island violation is much greater than the individual cost of any given island structure in isolation.

We can test the predictions of this simple reductionist theory with an acceptability judgment experiment that employs a factorial definition of island effects. Firstly, we can isolate the effect of dependency length on acceptability by contrasting a sentence with a short wh-dependency, an extraction from a matrix clause, (5a), with a sentence that contains a longer wh-dependency, an extraction from a embedded clause, (5b). Similarly, we can isolate the effect of processing island structures by contrasting a sentence with an island structure (5c) with a sentence that does not contain an island structure (5a). Finally, we can measure the effect on acceptability of processing both long-distance wh-dependencies and island structures -- the island effect itself -- by combining both in a single sentence (5d).

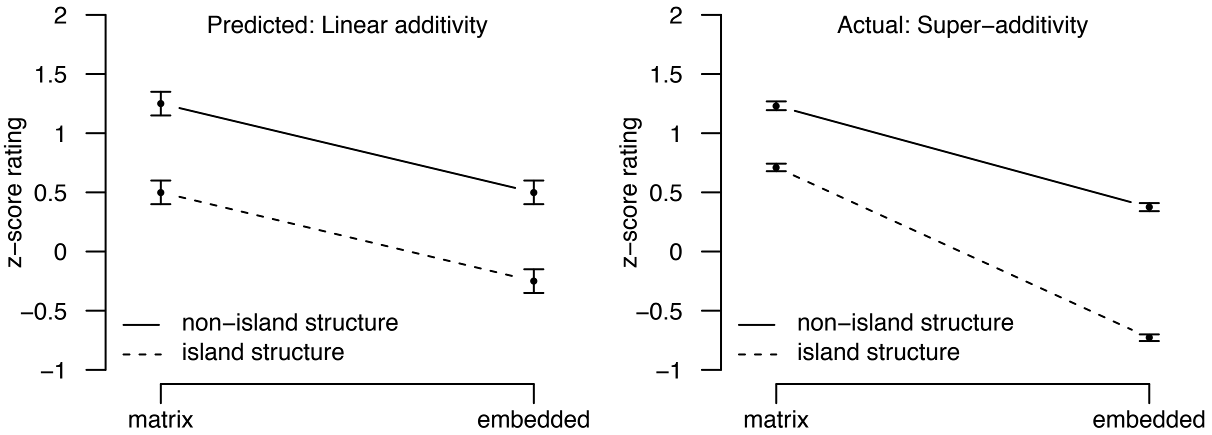
(5) A factorial design for measuring island effects: Structure x Gap Position

a. Who __ thinks that John bought a car? NON-ISLAND | MATRIX
b. What do you think that John bought __ ? NON-ISLAND | EMBEDDED
c. Who __ wonders whether John bought a car? ISLAND | MATRIX
d. What do you wonder whether John bought __ ? ISLAND | EMBEDDED

As the labels in (5) indicate, this design contains two factors (STRUCTURE and GAP-POSITION) each with two levels (ISLAND/NON-ISLAND and MATRIX/EMBEDDED) (see also Sprouse et al. 2011).

The simplest reductionist theory predicts that the relationship between the two processing costs should be linearly additive: the cost of processing long-distance dependences [(5a)-(5b)] plus the cost of processing whether clauses [(5a)-(5c)] should equal the cost of performing both together [(5a)-(5d)]. This prediction can be graphically represented using an interaction plot as in the left panel of Figure 1.

Figure 1: The left panel represents the prediction of the simplest reductionist theory. The right panel represents the actual results of using the factorial definition of *Whether island*s in (5) in an acceptability judgment experiment (see Section 5 for details of the experiment).



Crucially, a linearly additive relationship within a 2×2 factorial design results in parallel lines. Given the arrangement of conditions used in the left panel of Figure 1, the separation between the two lines reflects the main effect of whether clauses, and the slope of the lines reflects the main effect of long-distance dependencies. The rating of the island-violating sentence (condition (5d), which is in the bottom right quadrant of each panel of Figure 1) relative to the baseline condition (condition (5a), which is in the top left quadrant of each panel) is simply the sum of the two main effects. In this way, there is no need to invoke an additional syntactic constraint to explain the unacceptability of the island-violating sentence; the unacceptability is simply the result of (linearly) adding the two independently motivated costs together.

The problem with this simple reductionist theory is that island effects with wh-dependencies in English do not show linear additivity when tested using the factorial design in (5). The graph in the right panel of Figure 1 represents the results of a typical acceptability judgment experiment that employs the factorial definition of Whether islands (see Section 5 for details of the experiment). It is clear that the combined effect of the two costs is greater than the (linear) sum of the individual costs; in other words: [(5a)-(5b)] + [(5a)-(5c)] < [(5a)-(5d)]. This effect is thus a superadditive effect, since the overall difference is greater than the sum of the individual differences. A superadditive effect can be reflected statistically as an interaction, since the response to each level of one factor depends upon the level of the other. The same way that linear additivity can be visually identified by parallel lines, superadditivity can be visually by non-parallel lines.

2.2. AN ELABORATED RESOURCE-LIMITATION THEORY. Like the simple (linear) reductionist theory presented in the previous subsection, the resource-limitation theory of Kluender and Kutas (1993) builds on the observation that the sentences that give rise to island effects always contain two specific components: (i) a long-distance wh-dependency, and (ii) a complex syntactic structure. This means that the factorial definition in (5) is still appropriate for investigating the resource-limitation theory. However, the capacity based theory overcomes the limitation of the simple (linear) reductionist theory by including three additional assumptions that predict a superadditive interaction.

(6) Core assumptions of the resource-limitation theory (Kluender and Kutas 1993)

Component 1: There is a processing cost associated with the operations necessary to build long-distance wh-dependencies

Component 2: There is a processing cost associated with the operations necessary to build the syntactic structures that we call island structures

Linking hypothesis: Processing costs are reflected in acceptability judgments such that higher costs lead to lower acceptability

Additional assumptions of the resource-limitation theory

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

Limited Capacity: There is a limited pool of processing resources available that must be shared by all simultaneous processes

Overload: Additional unacceptability arises if the simultaneous processes necessary to complete the parse require more resources than are available in the limited pool.

Crucially, these three additional assumptions (Simultaneity, Limited Capacity, and Overload) predict that the combination of long-distance wh-dependencies and island structures should lead to the superadditive interaction seen in the right panel of Figure 1. In this way, the resource-limitation theory is an empirically adequate theory with respect to the pattern of acceptability judgments that arise in the factorial definition of island effects in (5).

2.3. EVALUATING REDUCTIONIST THEORIES. There are, in principle, three approaches to evaluating a reductionist theory. The first is to simply evaluate how well it captures the pattern of acceptability judgments. This is the approach that we took with the simple (linear) reductionist theory in Section 2.1, and the approach that we took in Section 2.2 with the resource-limitation theory. Whereas the simple (linear) theory cannot capture the pattern of acceptability that arises with the factorial definition of island effects, the more elaborated resource-limitation theory can. Crucially, grammatical approaches can capture this pattern equally well (by associating the ISLAND/EMBEDDED condition with a specific grammatical constraint that causes extremely low acceptability). This means that the basic pattern of average ratings does not help us to tease apart the resource-limitation theory and the resource-limitation theory.

The second approach to evaluating a reductionist theory is to directly assess the validity of each of the underlying assumptions. By definition, reductionist theories must rely upon mechanisms that are independently motivated: a reductionist theory is no longer reductionist if it must posit new mechanisms to account for the facts that it seeks to explain. Though a complete evaluation of the assumptions of the resource-limitation theory is beyond the scope of this paper, we offer brief comments about each mechanism/assumption.

Component 1: The processing cost associated with long-distance dependencies

The assumption of a processing cost associated with long-distance dependencies is well-motivated, based on various sources of evidence. Long distance dependencies are difficult to process for several reasons. The first is the fact that a displaced constituent, such as a wh-phrase, must be durably encoded in memory. The syntactic and semantic relationships that this wh-phrase enters into are often not resolved immediately, so either it needs to be actively maintained in working memory until it can be licensed and interpreted (Wanner & Maratsos, 1978); or it must be retrieved into active processing later in the sentence, a process which takes time and is potentially errorful (McElree, 2006). Therefore, there is a cost associated with an open dependency, whether this cost is cast as a maintenance cost or a retrieval cost. Wagers (2012) argues that, on balance, retrieval is the dominant contributor to the costliness of open long-distance dependencies. For the elaborated reductionist account to succeed, this is important given the observation that long-distance dependencies can span but not enter an island structure without greatly affecting acceptability (see the discussion of Simultaneity below). Thus, the overload in the resource-limitation theory cannot stem from maintenance alone.

The second reason that long distance dependencies challenge the processor stems from the fact that the tail of the dependency can be difficult for the parser to recognize (Fodor, 1978). In English, the grammatical domain containing the tail of the dependency is signaled by the absence of a constituent -- a gap. But, of course, constituents may be absent for reasons other than participation in a long-distance dependency. For example, some verbs, like read, are optionally transitive and correspondingly the absence of a pronounced DP after the verb read could correspond to its intransitive use, not the presence of a gap. In languages with resumption, like Irish or Bulgarian, the online processing of long-distance dependencies has hardly been studied (but see Pablos 2006 for Spanish topic-clitic dependencies). However the logical problem of recognizing the tail of the dependency persists. Pronouns used to mark the tail of long-distance dependencies in such languages are identical in form to pronouns used in anaphora (McCloskey, 2002). In short, there is not always a simple cue to where the tail of the dependency occurs.

The processor seems to respond to the dependency-tail recognition problem by successively hypothesizing gaps in all grammatically available positions in the clause (see review in Phillips & Wagers, 2007). The advantage of this strategy is that it leads the processor to close the dependency as soon as possible. The disadvantage is that sometimes a gap is hypothesized for a position that is filled by a pronounced DP, leading to a reanalysis (Crain & Fodor 1985, Stowe 1986).

In summary, long-distance dependency formation requires the engagement of the working memory system and recognizing the tail of a long-distance dependency is prone to misanalysis. Consequently Component I is a plausible contributor to complexity. The effect of long-distance wh-dependencies on acceptability is reliably present for all four island types tested below. These facts are discussed briefly in the results sections of each of the experiments.

Component 2: The processing cost associated with island structures

To our knowledge, the proposal that the construction of the island structures themselves always requires additional processing resources is only discussed within papers that deal directly with the resource-limitation theory (see Kluender and Kutas 1993 for ERP responses to island structures, and Hofmeister and Sag 2010 for reading time responses). However, it is not implausible that such a processing cost might exist, especially with island structures that also introduce complex semantic values (such as *Whether island*s). However, it should be noted that the experiments presented here do contain some evidence about the robustness of this assumption: the island-structure cost is only reliably present in the acceptability judgments of *Whether island*s, and is reliably absent in Complex NP islands and Subject islands. The unreliability of the island-structure cost raises a problem for the resource-limitation theory, as it is not clear how island effects could be an emergent property of a single processing cost. These facts are discussed briefly in the results sections of each of the experiments.

Linking hypothesis: Processing costs are reflected in acceptability judgments

The hypothesis that the ease of a recovering a grammatical representation affects the perception of its acceptability is well motivated and has been widely recognized in the field (see, e.g. the discussion in Chomsky 1965). Sentences containing center self-embeddings exemplify one extreme case where the difficulty of processing a sentence can depress its acceptability ratings. Likewise, the ease of misanalysing an ungrammatical sentence can raise its acceptability ratings (Gibson & Thomas 1999, Wagers, Lau and Phillips 2009). Temporary representations that arise in the course of incremental comprehension because of ambiguity or parsing strategies also affect acceptability (Fanselow and Frisch 2006, Sprouse 2008). However effects that can be securely attributed to processing difficulty are often relatively small (Sprouse 2008). How systematically such small modulations contribute to the acceptability contrasts between grammatical and ungrammatical sentences remains an interesting open question in the present context.

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

The simultaneity assumption is more complex than it first appears because there are in fact sentences that contain both an incomplete wh-dependency and an island structure simultaneously. For example, (7a) below contains a wh-dependency that spans a relative clause in subject position (a "double" island structure: both subjects and relative clauses are islands) and yet they have generally been considered to be acceptable in the linguistics literature. The island effect only arises when the wh-dependency terminates within the island structure as indicated by the underscore "gap" position in (7b).

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

To our knowledge, the simultaneity assumption of the resource-limitation theory has not been formulated in a fashion that captures the contrast in (7a-b). This is a potential problem. However, in the current discussion 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 an island structure.[2]

Limited Capacity: There is a limited pool of processing resources

The limited capacity assumption follows the widely held view that humans have limited working memory capacity (within the sentence processing literature see: King and Just 1991, Just and Carpenter 1992, Caplan and Waters 1999, Fedorenko et al 2006, 2007). It is a vigorously debated question what the underlying cause of this limitation is. For the purposes of the capacity theory it is not necessary that differences in working memory literally reflect differences in storage capacity (as if working memory were like a computer hard drive). For example, the limitation may reflect differences in the cognitive flexibility and executive control necessary to avoid or resolve retrieval interference (e.g. Braver, Gray and Burgess 2007, Kane, Conway, Hambrick and Engle 2007).

Overload: Additional unacceptability arises when the simultaneous processes necessary to complete the parse require more resources than are available in the limited pool.

[2] It should be noted that this assumption, though necessary to maintain some reductionist theories, would also be problematic for some of the evidence that has been presented to support those theories. The problem is that the necessary reformulation would likely focus on retrieval operations at the gap location rather than other processes at the boundaries of island structures. This could undermine the relevance of observations about how the parser behaves at the boundaries of island structures that have been presented as evidence for reductionist theories (e.g. Kluender and Kutas 1993b, Hofmeister and Sag 2010).

To our knowledge the overload linking hypothesis has only been proposed with respect to island effects, therefore we cannot yet evaluate it.

The third and final approach to evaluating reductionist theories is to examine predictions that made by the reductionist theory that are not shared by grammatical theories. This is the primary approach that we take in the present study. As previously noted, the resource-limitation and grammatical approaches differ with respect to the cause of the superadditive interaction that arises in acceptability judgment experiments that employ the factorial definition of island effects. Under the resource-limitation theory the superadditivity is due to an overload of limited resource capacity. Under grammatical theories the superadditivity is due to a grammatical constraint that specifically targets the ISLAND/EMBEDDED condition. In other words, the limited pool of resources plays a role in island effects for reductionist theories, but not for grammatical theories. This suggests that an experiment that tests the impact of variation in the limited pool of resources could be used to tease apart the two theories. The reductionist theory would predict a correlation between the size of the limited pool of resources and the size of the resulting island effects. The grammatical theory predicts no such correlation.

3. THE LOGIC OF THE PRESENT STUDIES.

Converting the basic observation that limited resource capacity plays a role in reductionist theories but not grammatical theories into a testable hypothesis is not as straightforward as it first appears. Crucially, the resource-limitation theory makes no predictions about the covariation of any single sentence type with individual differences in capacity. Rather the resource-limitation theory makes predictions about the relationship between resource capacity and the superadditive interaction that arises when four conditions are compared to each other using a factorial design. Therefore, any testable predictions should involve the relation between capacity differences and the superadditive interaction itself, rather than a relation between capacity differences and the acceptability of any individual sentence type. One plausible prediction of the resource-limitation theory is that the strength of the superadditive interaction that characterizes island effects should co-vary with the amount of available processing resources, as it is the limited quantity of processing resources that is responsible for the interaction in this approach. Specifically, under the resource-limitation theory individuals with larger processing resource capacity should exhibit weaker superadditive interactions (if any), and individuals with more limited processing resource capacity should demonstrate stronger superadditive interactions. Under the grammatical theory, processing resource capacity plays no role in the superadditive interaction, and therefore any individual differences in the strength of island effects that we might observe should not correlate with individual differences in processing resource capacity.

In order to investigate this prediction of the resource-limitation 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. We next discuss the rationale for the measures that we chose, as well as the specific statistical predictions of the resource-limitation and grammatical theories.

3.1. WORKING MEMORY AS A MEASURE OF PROCESSING RESOURCE CAPACITY.

There are a number of measures that are thought to reflect the capacity or efficiency 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). The primary concern in choosing working memory tasks for a correlation study is that a result that shows no correlation could either mean that there is no relationship between working memory capacity and island effects, or that we simply tested the wrong measure of working memory. In order to minimize this concern we chose our working memory tasks strategically. First, we chose a serial recall task, a simple span measure in which participants must report as many words as they can recall from a serially-presented list in the order they were presented. Simple span tasks can be distinguished from complex span tasks, in which the presentation of a to-be-recalled item and its recall are interrupted by another task, like reading a sentence or solving a math problem. It has been argued that complex span tasks are better predictors of linguistic performance (Daneman & Carpenter, 1980), because they presumably index both storage and processing efficiency. However, more recent research challenges the distinction between storage and processing and shows that even simple recall tasks can be a strong index of individual differences in working memory (Unsworth & Engle 2007). Moreover, there is a large component of shared variance between simple span and complex span tasks (Conway et al., 2005) and, correspondingly, performance on simple span tasks correlates with performance in language memory tasks (Roberts & Gibson 2002).

Second, we included the n-back task (Kirchner 1958, Kane & Engle 2002, Jaeggi et al 2008). In this task, participants are presented with a series of letters on a computer screen one at a time, and are asked to press a button if the letter currently on the screen was also presented n items previously. Each participant attempted three values of n: 2, 3, 4. This task has been shown to be generally uncorrelated with other simple and complex span tasks (Kane et al. 2007, Roberts & Gibson, 2002). Moreover Roberts and Gibson (2002) found that the n-back task correlated strongly with accuracy in a sentence memory task (recalling the subject or main verb of a clause), uniquely explaining 16% of the variance in performance. These findings suggest that (a) the n-back task indexes individual differences in the capacity and efficient of working memory; and (b) it captures different components of the working memory system than span tasks.

The two tasks we have chosen jointly correlate with most other popular WM tasks. As discussed in detail in Section 7, this does not eliminate the possibility that we did not test the correct component of the working memory system. However it does substantially decrease the likelihood of this error, as it is highly unlikely that any other extant working memory task would yield results that are significantly different than the results of the two tasks employed here.

3.2. MEASURING THE STRENGTH OF ISLAND EFFECTS. As discussed in Section 2, the crucial difference between resource-limitation and grammatical theories involves the source of the superadditive interaction elicited by the factorial design. Therefore we need a measure that captures the strength of the superadditive interaction, and can be compared to the working memory measures discussed above. A standard measure known as a differences-in-differences (DD) score achieves this (Maxwell and Delaney 2003). A DD score is calculated for a two-way interaction as follows. First, calculate the difference (D1) between the scores in two of the four conditions. To make the DD scores as intuitively meaningful as possible, we have defined D1 as the difference between the NON-ISLAND/EMBEDDED condition and the ISLAND/EMBEDDED condition. Second, calculate the difference (D2) between the scores in the other two conditions. For our purposes, D2 is the difference between the NON-ISLAND/MATRIX condition and the ISLAND/MATRIX condition. Finally, calculate the difference between these two difference scores.

Intuitively, the DD score measures how much greater the effect of an island structure is in a long-distance dependency sentence than in a sentence with a local dependency.

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

	D1 = (NON-ISLAND/EMBEDDED) − (ISLAND/EMBEDDED)	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 = (NON-ISLAND/MATRIX) − (ISLAND/MATRIX)	
	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 measure of the superadditive component of the interaction for each individual and for each type of island. Thus DD scores 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).

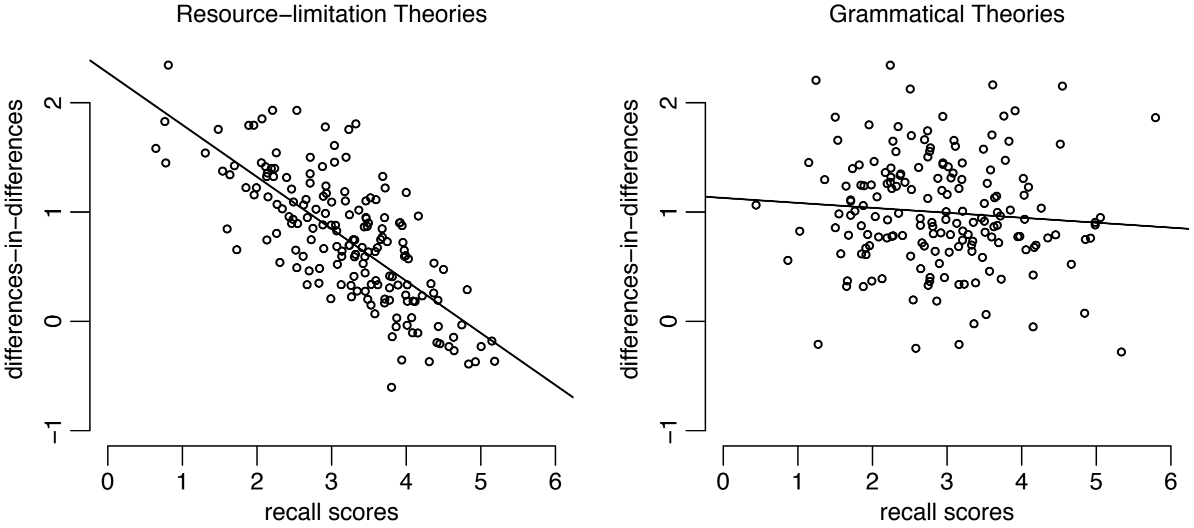
3.3. STATISTICAL PREDICTIONS OF THE RESOURCE-LIMITATION AND GRAMMATICAL THEORIES. The resource-limitation theory makes the prediction that if an individual has sufficient working memory capacity to handle both sets of processes simultaneously, then that individual's DD score (island strength) might be 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 resource-limitation theory believe that sufficient working memory capacity is indeed humanly possible. For example, Hofmeister and Sag (2010) make the following informal observations as potential support for the resource-limitation 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 2010:403]

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 plot DD scores as a function of working memory capacity for a sufficiently large sample of speakers, the resource-limitation theory predicts that the we should see a downward sloping trend as schematized in Figure 3a: as working memory scores increase, DD scores should decrease. Statistically speaking, the resource-limitation theory predicts that working memory capacity should be a significant predictor of DD scores, 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. Whether the actual relationship is linear is impossible to say without a more precisely specified model but the capacity theory at least predicts a monotone decreasing relationship. 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 resource-limitation (left-panel) and grammatical theories (right-panel)



In the sections that follow, we present two 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. Experiment 2 extended 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.

4. Experiment 1. In this section, we present the first of our two studies. The four island types investigated in our experiments were chosen because they are considered to be relatively mild 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). These milder island types are generally considered to be more readily interpretable, and therefore should be good candidates to display the variability in acceptability, and hence the correlation with working memory capacity predicted by reductionist theories.

4.1. MATERIALS AND METHODS.

PARTICIPANTS. 142 self-reported monolingual native speakers of English (76 female), all of them 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, such that related lexicalizations never appeared in the same list. Two pseudorandom orders of each of the 4 lists were created, such that items from 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.

(9) Whether islands
a. Who __ thinks that John bought a car? NON-ISLAND | MATRIX
b. What do you think that John bought __ ? NON-ISLAND | EMBEDDED
c. Who __ wonders whether John bought a car? ISLAND | MATRIX
d. What do you wonder whether John bought __ ? ISLAND | EMBEDDED

(10) Complex NP islands
a. Who __ claimed that John bought a car?
b. What did you claim that John bought __?
c. Who __ made the claim that John bought a car?
d. What did you make the claim that John bought __?

(10) Subject islands
a. Who __ thinks the speech interrupted the TV show?
b. What do you think __ interrupted the TV show?
c. Who __ thinks the speech about global warming interrupted the TV show?
d. What do you think the speech about __ interrupted the TV show?

(11) Adjunct islands
a. Who __ thinks that John left his briefcase at the office?
b. What do you think that John left __ at the office?
c. Who __ laughs if John leaves his briefcase at the office?
d. 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." 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. Though we believe that the z-score transformation is the most appropriate method for analyzing scale-based acceptability judgment data, it should be noted that we also ran all of the regression analyses reported using the raw scores rather than the z-score transformed scores with no change in results (see Section 6.1). The means and standard deviations for each condition are reported in Table 1.

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

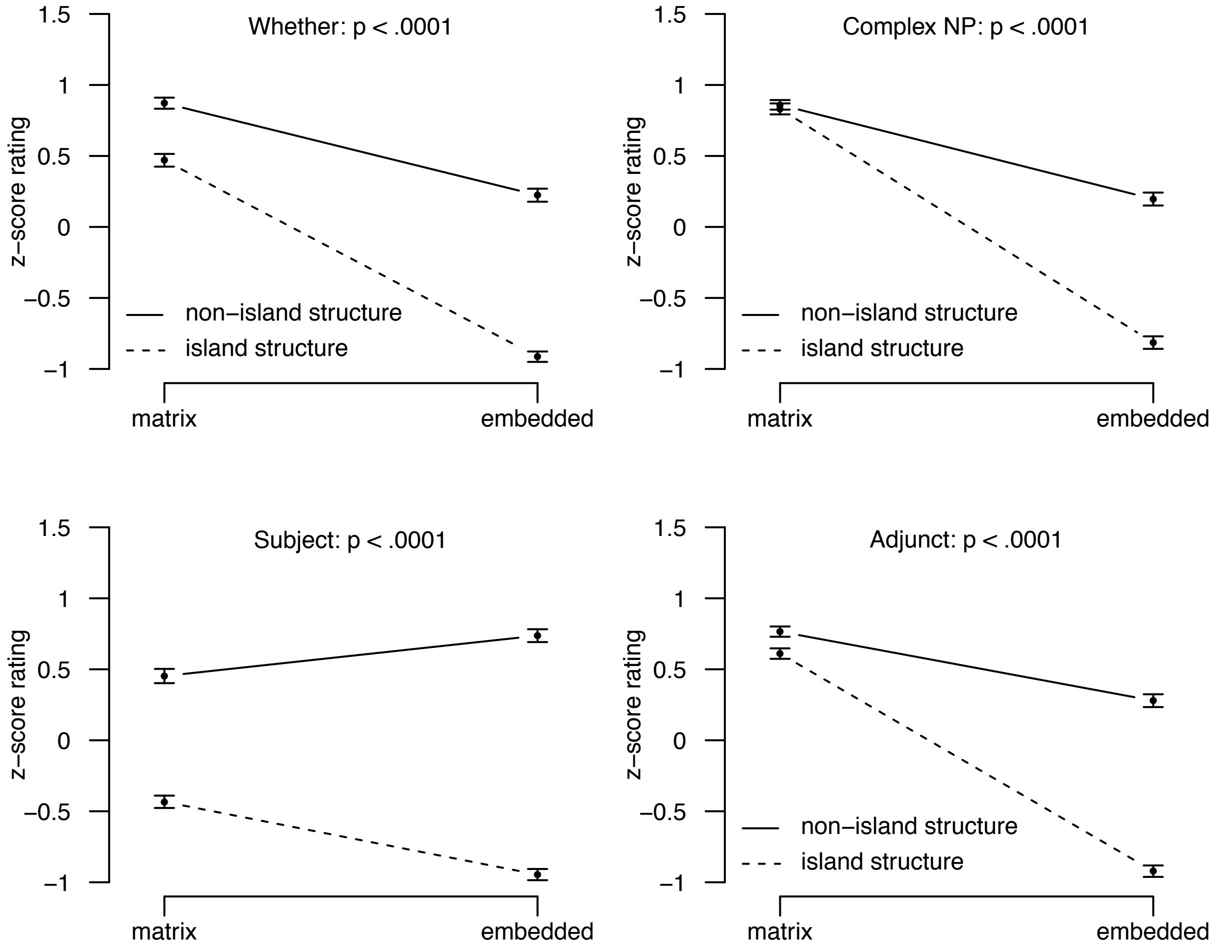
	Whether	Complex NP	Subject	Adjunct
NON-ISLAND | MATRIX	0.87 (0.60)	0.86 (0.58)	0.45 (0.86)	0.77 (0.64)
NON-ISLAND | EMBEDDED	0.22 (0.76)	0.20 (0.78)	0.74 (0.78)	0.28 (0.81)
ISLAND | MATRIX	0.47 (0.69)	0.83 (0.61)	-0.43 (0.73)	0.61 (0.65)
ISLAND | EMBEDDED	-0.91 (0.60)	-0.81 (0.65)	-0.95 (0.61)	-0.92 (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 constructed linear mixed effects models with items and participants included as random factors on each of the island types using GAP-POSITION 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). We also performed pairwise comparisons on the two non-island conditions, to test for an independent effect of length;- and on the two matrix gap conditions, to test for an independent effect of structure. Table 2 reports the p-values for each factor and the interaction of the full 2×2 model (upper panel) and the two pairwise comparisons (lower panel).

Table 2: Experiment 1, *p*-values for the two-way linear mixed effects models for each island type (n=142)

	Whether	Complex NP	Subject	Adjunct
FULL 2 × 2 MODEL				
Main effect of GAP-POSITION	.0001	.0001	.0144	.0001
Main effect of STRUCTURE	.0001	.0001	.0001	.0001
GAP-POSITION x STRUCTURE	.0001	.0001	.0001	.0001
PAIRWISE COMPARISONS				
GAP-POSITION (STRUCTURE=NON-ISLAND)	.0001	.0001	.0001	.0001
STRUCTURE (GAP-POSITION=MATRIX)	.0001	.5694	.0001	.0564

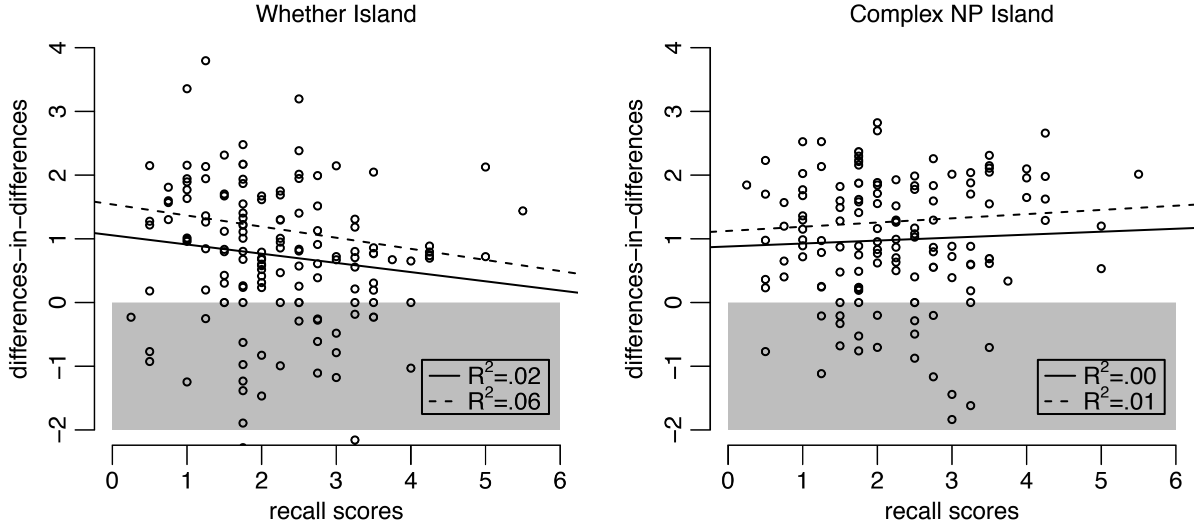
Figure 3: Experiment 1, interaction plots for each island type (n=142)



There was a significant main effect of GAP-POSITION and STRUCTURE for each island type. Crucially, there was a significant interaction of GAP-POSITION and STRUCTURE for every island type (p < .0001). However, the pattern of results for Complex NP islands was not as predicted by the resource-limitation theory: there was no independent cost of the island structure. For Adjunct islands, the pairwise comparison on STRUCTURE reached only marginal significance. This pattern of results presents a problem for the generality of the resource-limitation theory, as one of the fundamental processing costs did not appear to be robust in all of the island types. 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 a significant effect of island 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 (8) and plotted as a function of serial recall scores in Figure 5. 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 or equal to 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-additive interactions, which raises questions about how to interpret participants who present sub-additive island effects. One possibility is that DD scores below 0 may reflect a type of noise that we may not want to influence the linear regression. If they are indeed noise, then eliminating these scores from the analysis should increase the likelihood of finding a significant correlation in the data. On the other hand, it is possible that these DD scores represent participants who truly do not perceive a classic superadditive island effect. In this case, including these scores should increase the likelihood of finding a significant correlation in the data. Because we have no firm basis for choosing between these two possibilities, we decided to report both analyses. The removal procedure in the second analysis affected 27 participants for Whether islands (19%), 20 participants for Complex NP islands (14.1%), 19 participants for Subject islands (13.4%), and 16 participants for Adjunct islands (11.3%). Table 3 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.



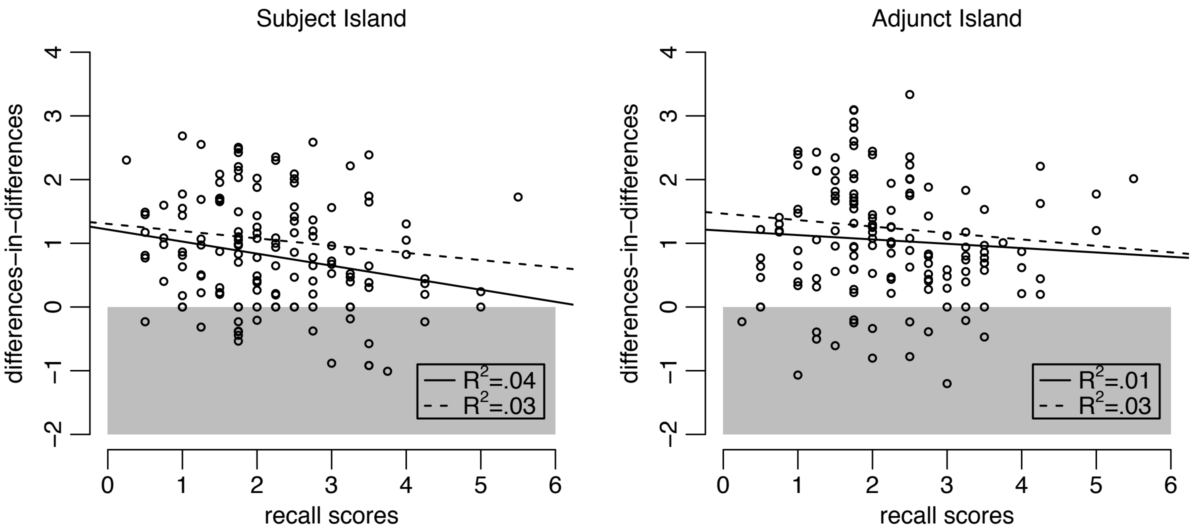


Table 3: Experiment 1, linear regression modeling differences-in-differences scores as a function of serial recall scores (n=142)

	island	line-of-best-fit		goodness-of-fit	significance test	
		intercept	slope	R^2	*t*-statistic	*p*-value
All DDs	Whether	1.05	-0.14	.02	-1.65	.10
	Complex NP	0.88	0.05	.00	0.58	.56
	Subject	1.22	-0.19	.04	-2.42	.02
	Adjunct	1.20	-0.07	.01	-0.92	.36
DDs ≥ 0	Whether	1.54	-0.16	.06	-2.52	.01
	Complex NP	1.16	0.05	.01	0.81	.42
	Subject	1.34	-0.13	.03	-1.98	.05
	Adjunct	1.53	-0.13	.03	-1.97	.05

A simple linear regression analysis finds the line that minimizes the distance between all of the points and the line itself by using a least-squares measure. The line-of-best-fit is the line that minimizes the sum-of-square differences between the actual data and the line. Three properties of the line are relevant for our analysis. The first is the mathematical formula of the line itself: the y-intercept and the slope associated with a one-unit change in the predictor variable (in Table 3 the predictor variable is recall score). The second property is how well the line explains the data, the goodness-of-fit of the line. Goodness-of-fit is critical to the interpretation of the results of a linear regression because the least-squares fitting procedure always returns a line-of-best-fit, even when that line does not fit the data well at all. The R^2 statistic provides an intuitive measure of the proportion of the variance of the data captured by the line (between 0 and 1). R^2 is calculated as follows: first, calculate the sum-of-squares between each of the data points and a horizontal line that passes through the mean of the y-values (SS_{total}); second, calculate the sum-of-squares between each of the data points and the line-of-best-fit (SS_{error}); next, divide SS_{error} by SS_{total} to determine the proportion of variance that is unexplained by the line-of-best-fit; finally, subtract this value from 1 to determine the proportion of variance that is explained by the line-of-best-fit. The third property returned by the linear regression analysis is a null hypothesis statistical significance test that answers the following question: Assuming that the underlying relationship between x and y should result in a slope of 0 (i.e. assuming the null hypothesis), what is the probability of finding the observed slope in a given data sample?

Table 3 reports the results of the linear regressions: line-of-best-fit (intercept and slope), goodness-of-fit (R^2), and significance of the slope (t-statistic and p-value). In the parameters of the lines-of-best-fit we find that 3 out of 4 slopes are negative, as predicted by the resource-limitation theory. However, the line-of-best-fit for the Complex NP island has a positive slope, which is not predicted by either theory. The goodness of fit of the line-of-best-fit is poor for all 4 island types. The best possible model was one of the biased models, and it only captured 6% of the variance in its data set, whereas 4 of the models captured 0-2% of the variance. 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 resource-limitation theory has an R^2 of .50 (i.e. 50% of the variance in the data is explained by the line). This figure is more comparable to significant correlations found in the psycholinguistics literature, such as the sentence memory findings of Roberts & Gibson (2002). 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 potentially noisy DD scores) are not at all what one would predict under a theory like the resource-limitation theory, which relies heavily on a single factor for its explanatory power. These goodness-of-fit results indicate that the expectations of the resource-limitation theory are not met in the relationship between DD scores and recall scores.

Finally, we can examine the results of the null hypothesis significance test of the slopes of the lines. Recall that this particular test reports the probability of observing the best-fit slope if the true population slope were 0 (i.e. a horizontal line). Four of the eight regressions returned lines with slopes that were significantly different from 0 at a level of $p < .05$, with three of those in second set of regressions (DDs greater than or equal to 0). Unfortunately, because the goodness-of-fit of the lines is so low, these results are not particularly meaningful. 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 Experiment 2 there were only two lines with slopes that were unlikely given the null hypothesis, and both of those had a positive slope (in contrast to the significant slopes in Experiment 1, which were all negative). These conflicting results underscore the fact that the slopes of poorly fitting trend lines provide scant evidence of an interpretable relationship between recall and the DD scores.

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. To that end, 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, a second 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 (see also Section 3 and Section 7). 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 capture distinct variance 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 used in Experiment 1 is more like the set of contrasts considered in the theoretical syntax literature, but crucially it led to a much smaller interaction than the standard STRUCTURE x GAP-POSITION 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 × GAP-POSITION 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, Sprouse 2011) 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.

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:

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

As in psychophysical ME, the standard in syntactic ME remains visible throughout the experiment. Magnitude estimation has increasingly gained currency in experimental linguistics research (Keller 2000, 2003, Featherston 2005a, 2005b, Sprouse 2009), although recent evidence suggests that ME does not result in more sensitive judgment data (Sprouse 2011, Weskott and Fanselow 2011). Nevertheless, ME seems well suited to the present study because of its potential to capture a wider range of ratings using the unbounded positive number line.

5.1. MATERIALS AND METHODS.

PARTICIPANTS. 176 self-reported monolingual native speakers of English (152 Female), all University of California Irvine undergraduates, participated in this experiment for either course credit or $5. 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: Whether islands, Complex NP islands, Subject islands, and Adjunct island. For each type of island, extraction site and structural environment was manipulated in a 2×2 design, as discussed in Section 2, yielding a total of 16 critical conditions in the experiment. 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.

(12) 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 first 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 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 & 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 the proportion of hits, i.e. correct responses to n-back target-present trials, and the proportion of false alarms, i.e. 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. 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 (though it should be noted again that we performed the same analyses on the raw data and obtained the same results; see Section 6.1). DD scores were calculated using the formula presented in Section 3.

Table 4: Experiment 2, means and standard deviations of z-scored magnitude estimates for each condition (n=173)

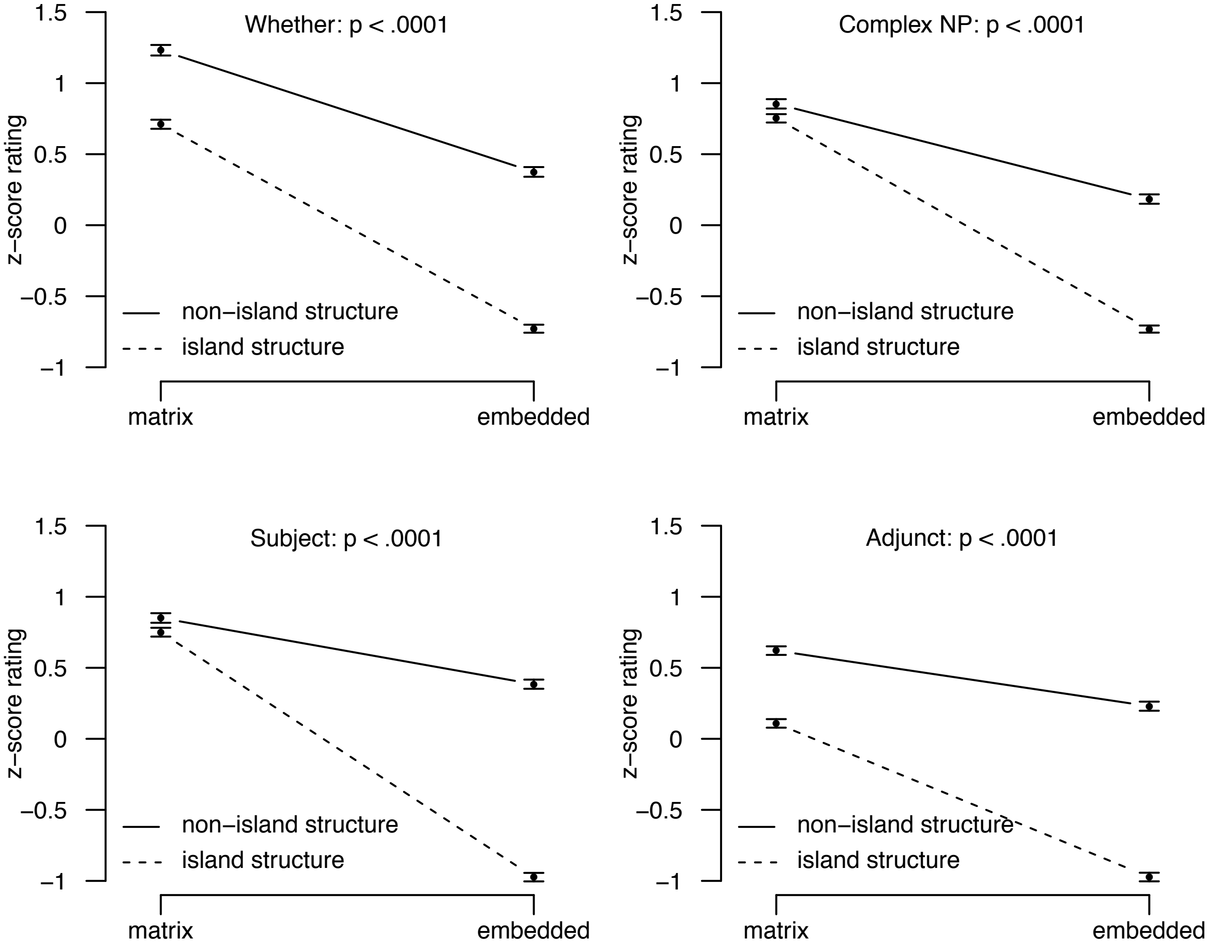
	Whether	Complex NP	Subject	Adjunct
NON-ISLAND | MATRIX	1.23 (0.74)	0.86 (0.76)	0.85 (0.77)	0.62 (0.80)
NON-ISLAND | EMBEDDED	0.38 (0.72)	0.18 (0.82)	0.38 (0.83)	0.23 (0.79)
ISLAND | MATRIX	0.71 (0.67)	0.75 (0.71)	0.75 (0.79)	0.11 (0.81)
ISLAND | EMBEDDED	-0.73 (0.63)	-0.73 (0.57)	-0.97 (0.61)	-0.97 (0.72)

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 revealed a significant main effect of GAP-POSITION and STRUCTURE for each island type. However, because the interactions were super-additive, it is possible that the ISLAND | EMBEDDED 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 NON-ISLAND | MATRIX and NON-ISLAND | EMBEDDED conditions. The structure cost was isolated with a pairwise comparison of the NON-ISLAND | MATRIX and ISLAND | MATRIX conditions. As Table 5 indicates, the effect of GAP-POSITION was significant for every island type as expected. However, the effect of STRUCTURE was not significant for complex NP and Subject islands. 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 was only reliably present in the Whether island, and was reliably absent in the Complex NP island and the corrected Subject island design.

Table 5: Experiment 2, *p*-values for the 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
FULL 2 × 2 MODEL				
Main effect of GAP-POSITION	.0001	.0001	.0001	.0001
Main effect of STRUCTURE	.0001	.0001	.0001	.0001
GAP-POSITION X STRUCTURE	.0001	.0001	.0001	.0001
PAIRWISE COMPARISONS				
GAP-POSITION (STRUCTURE=NON-ISLAND)	.0001	.0001	.0001	.0018
STRUCTURE (GAP-POSITION=MATRIX)	.0001	.2142	.3335	.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. As before we also ran all analyses using the raw ratings with no change in the results (see Section 6.1). 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 greater than or equal to zero. The exclusion of DD scores below zero affected 36 scores for Whether islands (20.8%), 27 for CNPC islands (15.6%), 17 for Subject islands (9.8%), and 31 for Adjunct islands (17.9%).

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.

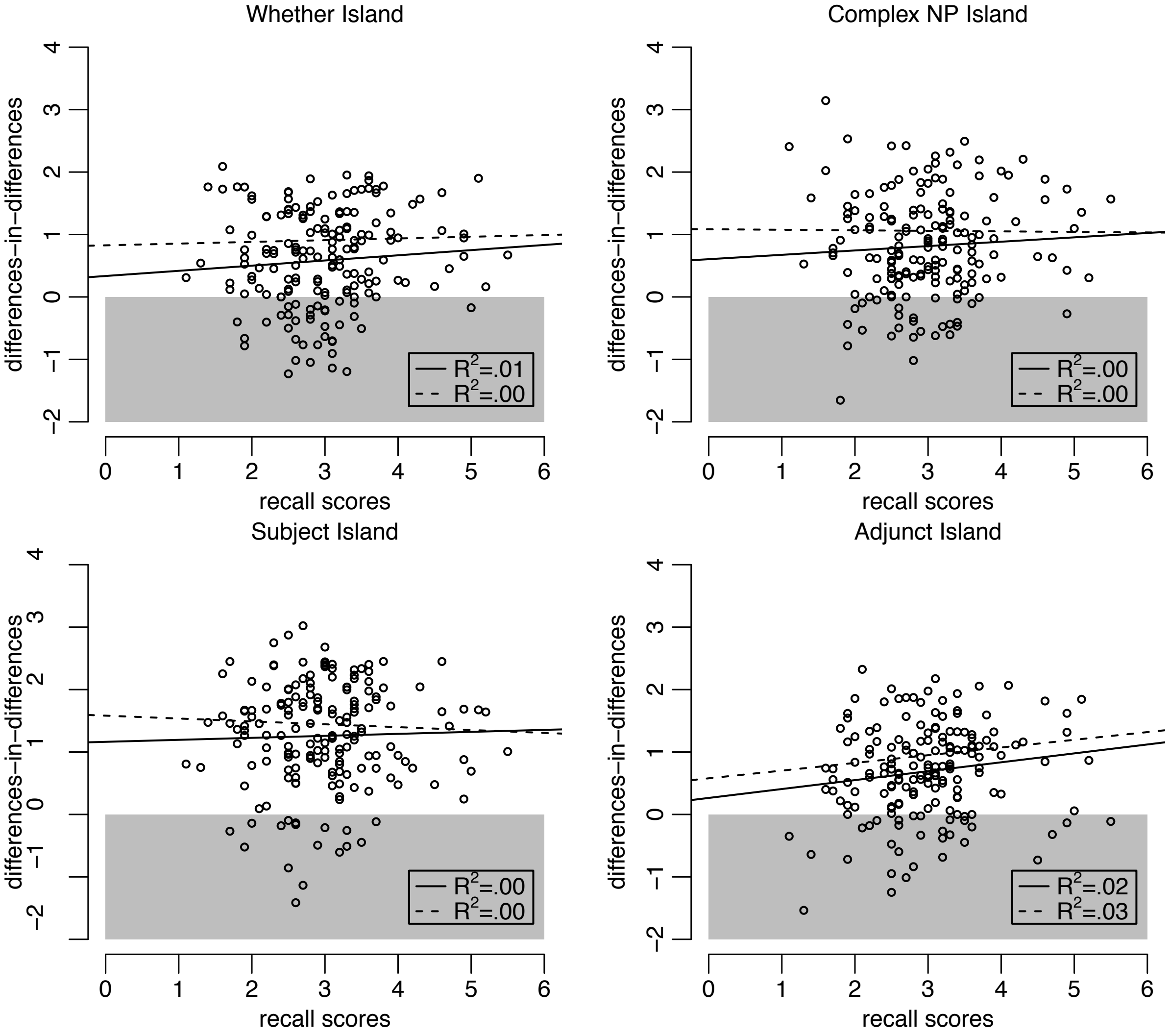


Table 6: Experiment 2, linear regression modeling differences-in-differences scores as a function of serial recall scores (n=173)

	island	line-of-best-fit		goodness-of-fit	significance test	
		intercept	slope	R^2	*t*-statistic	*p*-value
All DDs	Whether	0.34	0.08	.01	1.05	.29
	Complex NP	0.60	0.07	.00	0.88	.38
	Subject	1.16	0.03	.00	0.39	.70
	Adjunct	0.26	0.14	.02	2.02	.04
DDs ≥ 0	Whether	0.83	0.03	.00	0.48	.64
	Complex NP	1.04	0.00	.00	0.01	.99
	Subject	1.58	-0.05	.00	-0.71	.48
	Adjunct	0.58	0.12	.03	2.01	.05

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. First of all, seven out of the eight lines-of-best-fit have positive slopes (only Subject islands in the second set of regressions yielded a negative slope), contrary to the predictions of the resource-limitation theory. More importantly, all of the R^2 values are extremely low, even lower than Experiment 1 – five are zero, and the other three are .01, .02, and .03. This suggests that none of the lines are particularly meaningful models of the data. The extremely low R^2 values make an analysis of the significance tests unnecessary; however, only Adjunct islands revealed slopes that were significantly unlikely assuming the null hypothesis.

Differences-in-differences as a function of n-back. The descriptive results of the n-back tasks are reported in Table 7.

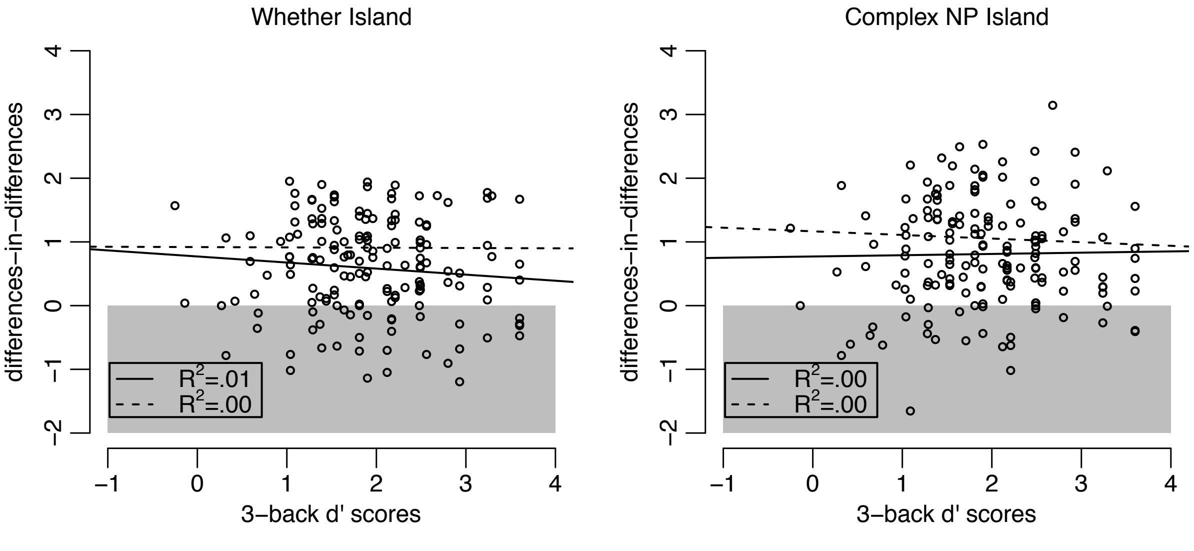
Table 7: Experiment 2, means and standard deviations for the *n*-back tasks (n=161)

	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 simple 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 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 (e.g. a large 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. To do so we performed a Principal Components Analysis (PCA) on the factors prior to the regression (Pearson 1901, Jolliffe 2002). Because there were 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 allows us to include them in the model without violating the assumptions of multiple linear regression. Finally, the multi-dimensionality of multiple linear regression tends to inflate R^2 values. As such, we report adjusted R^2 values for multiple linear regressions, which can be intuitively thought of as R^2 values minus an adjustment factor that corrects for the inflation inherent to multiple linear regression. Because of this subtraction, it is possible for adjusted R^2 values to be below zero in cases where the inflated R^2 was at or near zero.

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.



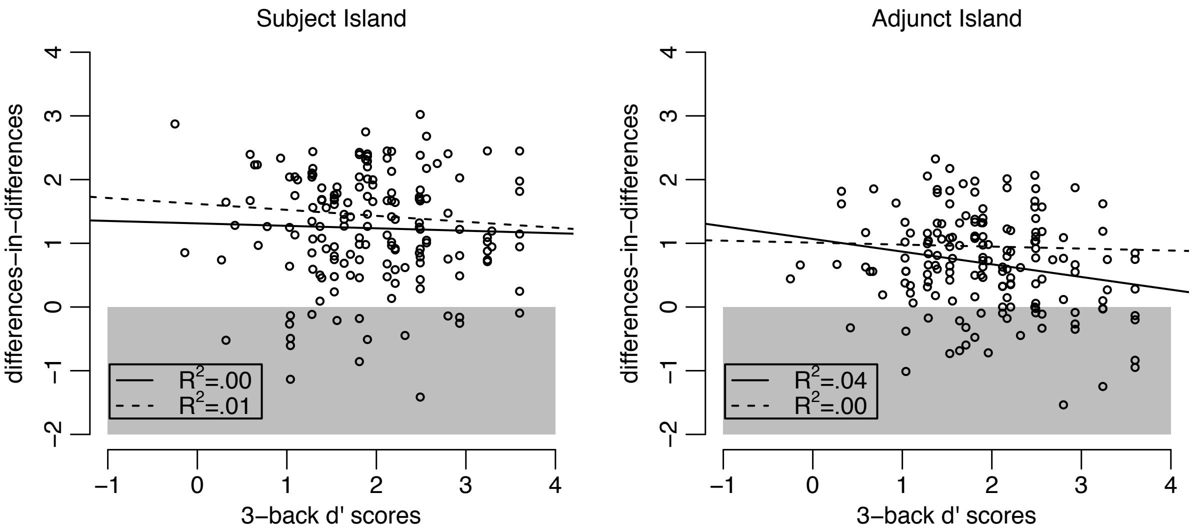


Table 8: Experiment 2, linear regression modeling differences-in-differences scores as a function of 3-back scores (n=161)

	island	line-of-best-fit		goodness-of-fit	significance test	
		intercept	slope	R^2	*t*-statistic	*p*-value
All DDs	Whether	0.77	-0.10	.01	-1.19	.24
	Complex NP	0.77	0.02	.00	0.24	.81
	Subject	1.31	-0.04	.00	-0.44	.66
	Adjunct	1.07	-0.20	.04	-2.72	.01
DDs ≥ 0	Whether	0.92	0.00	.00	-0.07	.94
	Complex NP	1.16	-0.06	.00	-0.81	.42
	Subject	1.62	-0.09	.01	-1.31	.19
	Adjunct	1.01	-0.03	.00	-0.48	.63

The results of the linear regression on 3-back scores is similar to the results of the recall scores: five of the models have adjusted R^2 values that are at 0, and the other two are .01 and .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 substantive change in the results.

For the model including all three n-back scores, we included all three of the orthogonal components returned by the PCA in the model because every component accounted for a sizeable portion of the variance in the original data set (making it impossible to reasonably eliminate any of the components): 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 resource-limitation theory. However, even with all three n-back scores included in the model, none of the models captured a meaningful amount of the variance, as all of the adjusted R^2 values were at or near zero (see the Appendix for the complete statistical details).

Table 9: 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 ≥ 0	-.02	.01	-.01	-.02

COMBINING BOTH SERIAL RECALL AND N-BACK SCORES. 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 percentages of original variance respectively: 33%, 28%, 20%, and 19%.

Table 10: 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 ≥ 0	-.02	.00	-.01	.00

As Table 10 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.

5.3. CONCLUSION. 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 parallels between the results of Experiment 1 and Experiment 2 suggests that the independence of working memory capacity and island judgments is not an artifact of 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. ADDITIONAL ANALYSES. Though we believe that the analyses presented in the previous subsections are the most appropriate, and therefore the most likely to reveal a correlation between working memory capacity and island effects, in this section we describe results of additional analyses that address possible concerns with the choices that we made in our primary analyses. For space reasons we cannot report every possible analysis of this data. Therefore we have also made the original data set available on the first author's website so that interested readers may verify these results for themselves.

6.1. RAW RESPONSES RATHER THAN Z-SCORE TRANSFORMED RESPONSES. Instead of z-score transforming responses prior to analysis (to remove potential scale differences between participants), one could perform the analyses directly on the raw ratings. Table 11 reports the R^2 values for simple linear regression using serial recall and multiple linear regression using the principal components of the three n-back tasks. As in the previous analyses on transformed scores, neither recall nor *n*-back scores account for more than a few percent of the variance.

Table 11: Experiment 2, R^2 and adjusted R^2 values for linear regressions that use the raw acceptability ratings.

		Whether	Complex NP	Subject	Adjunct
serial recall (R^2)	All DDs	.00	.01	.01	.03
	DDs ≥ 0	.00	.00	.00	.03
n-back (adjusted R^2)	All DDs	.00	.00	-.01	.02
	DDs ≥ 0	-.02	.00	.00	.01

6.2. ELIMINATING FATIGUE AS A POTENTIAL CONFOUND. Another possible concern is that the length of the acceptability judgment task may have led to fatigue during the later ratings, potentially obscuring a relationship between working memory capacity and island effects. The resampling simulations presented in Section 6 likely control for this possibility, but it is also straightforward to run a direct analysis: Table 12 reports the R^2 values that result from linear regressions that use only the first rating of each condition. The results are similar to previous analyses.

Table 12: Experiment 2, R^2 and adjusted R^2 values for linear regressions that only use the first acceptability rating of each condition.

		Whether	Complex NP	Subject	Adjunct
serial recall (R^2)	All DDs	.00	.01	.00	.00
	DDs ≥ 0	.01	.00	.00	.00
n-back (adjusted R^2)	All DDs	-.02	-.01	.00	.01
	DDs ≥ 0	-.02	-.01	-.01	.01

6.3. RANDOMIZATION-BASED SIGNIFICANCE TESTS. The analyses presented in Sections 4 and 5 primarily relied on the intuitive interpretation of R^2 values as the proportion of variance accounted for by the line-of-best-fit, and our scientific judgment that such small R^2 values are unlikely to arise if there truly were a significant relationship between working memory capacity and island effects. One way of formalizing this intuition is to run simulations to estimate the distribution of results that would arise if the null hypothesis were true. In other words, we can use the data from this experiment to simulate what would happen if there really were no relationship between working memory capacity and island effects. Then we can compare the actual results of our experiments to these simulated results. In the statistics literature this is known as a (null hypothesis) bootstrap test (Edgington and Onghena 2007).

The first step in running a bootstrap test is to choose the statistic that we are going to test. Up to this point, we have primarily focused on R^2 values because they are a straightforward measure of the goodness-of-fit of the line-of-best-fit. Unfortunately, R^2 values are not an ideal choice for simulation based analyses like the bootstrap test because they do not report the direction of the correlation line. This means that lines-of-best-fit that indicate a positive correlation between memory and island effects (the opposite of the predicted direction under the resource-limitation theory) would be collapsed together with lines-of-best-fit that indicated a negative correlation (the actual prediction of the resource-limitation theory). We could report R^2 values in the analyses in previous sections because we also reported the slope of the lines in the tables and graphs. That is not possible in the simulations.

For the bootstrap simulations we instead focus on Pearson's r. Pearson's r, or the coefficient of correlation, is a measure of the strength and direction of a correlation between two variables. Pearson's r ranges in value between -1, which indicates a perfect negative correlation, and 1, which indicates a perfect positive correlation. An r-value of 0 indicates no correlation between the variables whatsoever. In the least-squares regressions we have reported, R^2 is simply the square of the coefficient of correlation. Given that the resource-limitation theory predicts a negative correlation between memory and island effects, one should expect that the Pearson's r for the correlation between memory and island effects would be negative, and relatively large. Like R^2 values, there are no objective criteria for deciding that r is "large", though Pearson suggested that an r of .1 be considered a weak correlation, .3 be considered a medium correlation, and .5 be considered a strong correlation. Because we estimate the distribution of r values with a bootstrap test, we can instead use the distribution itself to estimate the probability of the actual r values under the null hypothesis.

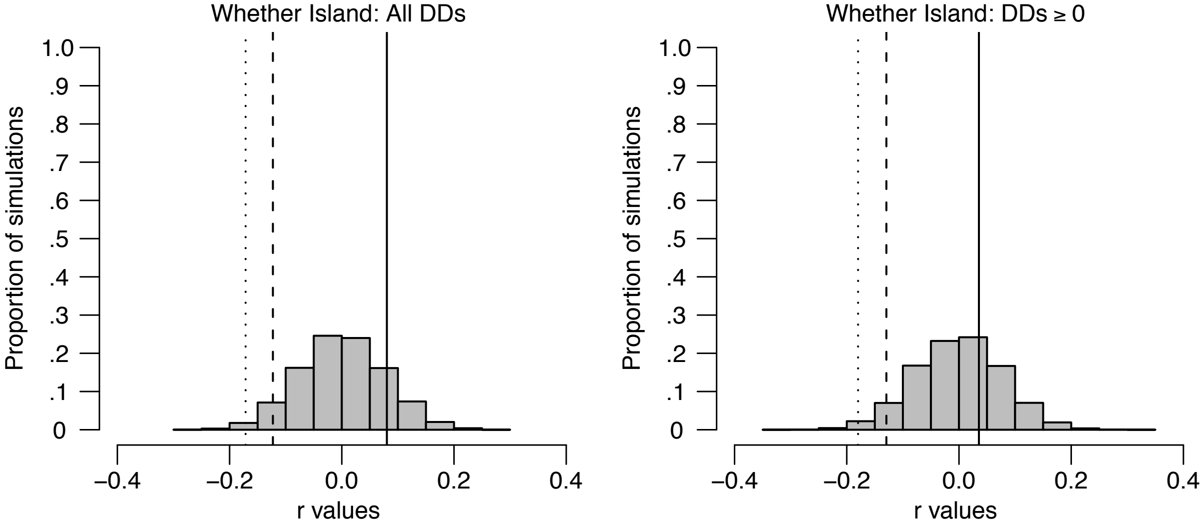
The algorithm for the bootstrap test is as follows:

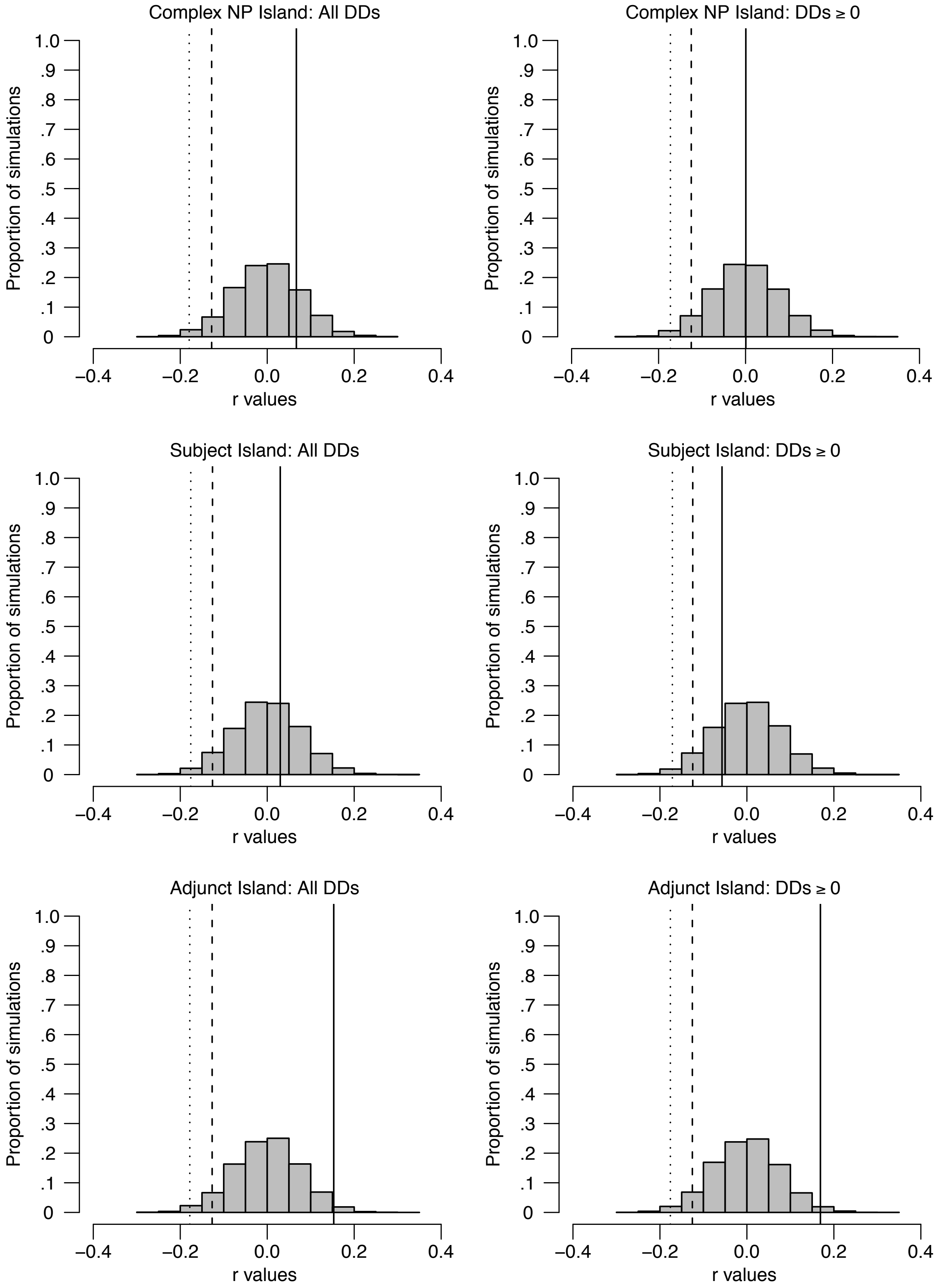
1. Randomly select 173 DD scores from the set of DD scores that we obtained for the island of interest. We allow the selection process to choose each DD score multiple times if necessary (as many times as it arises in the selection process). This is called sampling with replacement.
2. Pair each of the 173 randomly selected DD scores with the 173 memory capacity scores from the original experiment. Because the selection process in step 1 was random, this pairing is also random. The random nature of this pairing is exactly what is predicted by the null hypothesis: no relationship between memory scores and DD scores.
3. Calculate Pearson's r for the correlation between the memory scores and randomly selected DD scores.
4. Record the results.

5. Repeat the process (steps 1-4) 10,000 times to derive a distribution of linear regression results for 10,000 randomly selected pairings. This approximates the distribution of results that are expected under the null hypothesis.
6. Compare the actual r value that was obtained in the experiment to the simulated distribution. As with all null hypothesis tests, if it is relatively unlikely given the simulated distribution (i.e. fewer than 5% of simulated r values are more negative than the actual r value), we can conclude that there is evidence that the null hypothesis is unlikely.
7. Repeat this process (steps 1-6) for each island type.

The bootstrap algorithm sketched above provides us with two useful pieces of information. First, it provides an estimate for the distribution of r values that we should expect under the null hypothesis (i.e. no relationship between memory capacity and island effects) for the range of scores that we obtained in these experiments. This frees us from some of the subjectivity involved in defining r values as large or small. Second, it allows us to estimate the probability of obtaining r values that are more negative than the actual r value for each island; in other words, an empirically derived *p*-value.

Figure 9: Results of the bootstrap test for correlation (Pearson's r) between DD scores and recall scores in Experiment 2 (10,000 simulations). The solid line indicates the observed r value. Values to the left of the dashed line would be significant at *p*<.05. Values to the left of the dotted line would be significant at *p*<.01.




Complex NP Island: All DDs
Complex NP Island: DDs ≥ 0
Subject Island: All DDs
Subject Island: DDs ≥ 0
Adjunct Island: All DDs
Adjunct Island: DDs ≥ 0
Proportion of simulations
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
r values
−0.4
−0.2
0.0
0.2
0.4

As Figure 9 makes clear, there were no significant correlations between DD scores and recall scores in the direction predicted by resource-limitation theories (i.e. negative r values) for any of the island types. These randomization-based results corroborate the results from Section 5 that there is no substantial relationship between acceptability judgments and memory resources.

6.4. SIMULATING THE DISTRIBUTION OF POSSIBLE CORRELATION COEFFICIENTS IN THE DATA SET. 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, we averaged those four ratings into a single score to perform the regression. In the process of averaging those four trials we eliminated some of the uncertainty in the rating process. It is possible that some readers may be concerned this averaging procedure could have obscured a relationship between working memory capacity and island effects in the data that we collected.

To address this concern, we performed a second series of resampling simulations using the data from Experiment 2. Resampling simulations allow us to treat the sample of data that we obtain in a laboratory experiment as if it were the complete population of data. We can then simulate the results of running thousands of experiments on this population of data. These simulations provide us with two useful pieces of information: (i) an estimate of the range of possible results that could be obtained if one were to actually re-run the experiment, and (ii) an estimate of the probability of each of the possible results. For our purposes, these resampling simulations can provide an estimate of the range of possible r values for each island type, as well as the probability of each r value.

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.

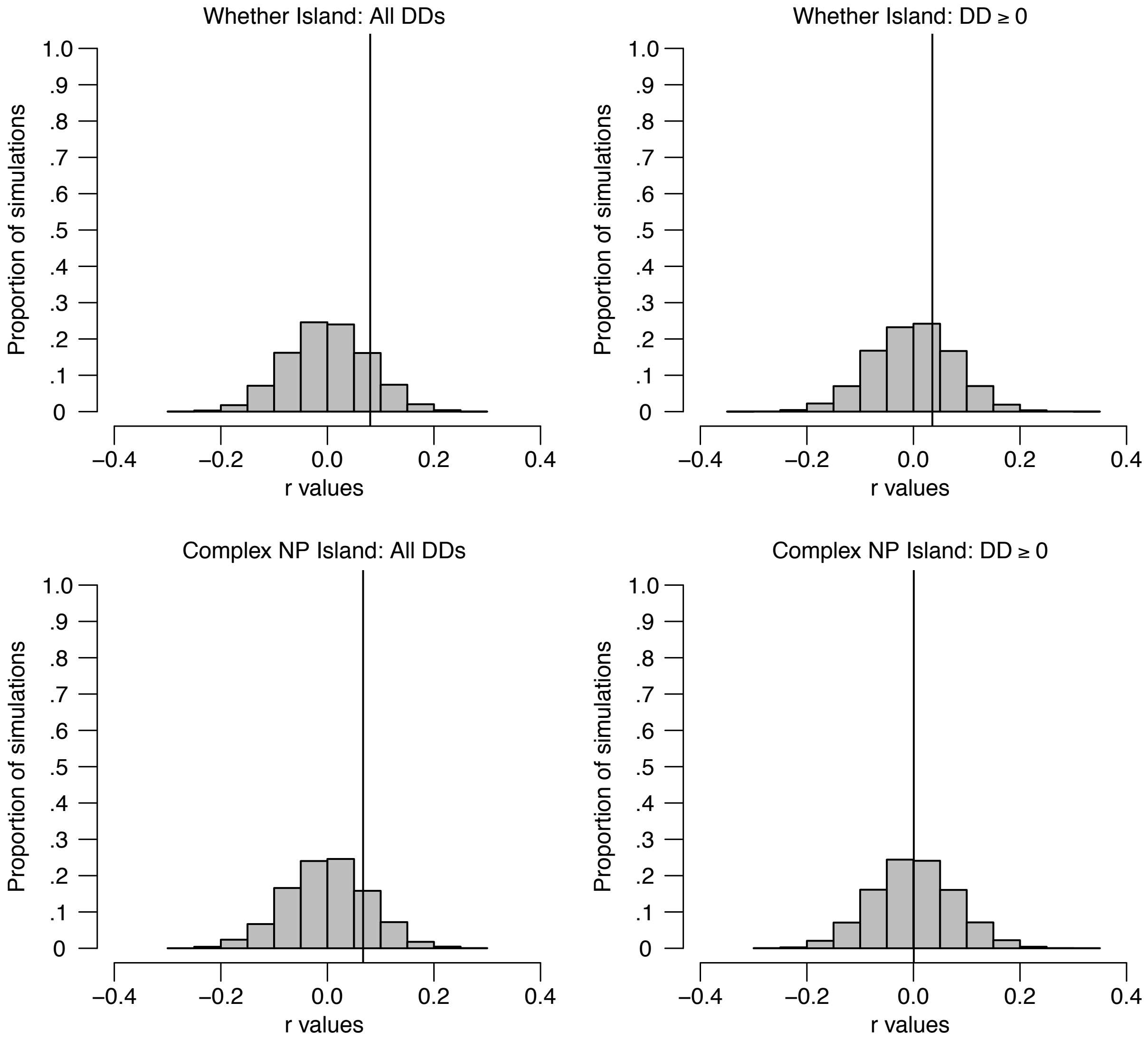
4. Next we calculated a DD score for each island type for each participant using the 10,000 scores we obtained in Step 3. This resulted in 10,000 DD scores for each participant.

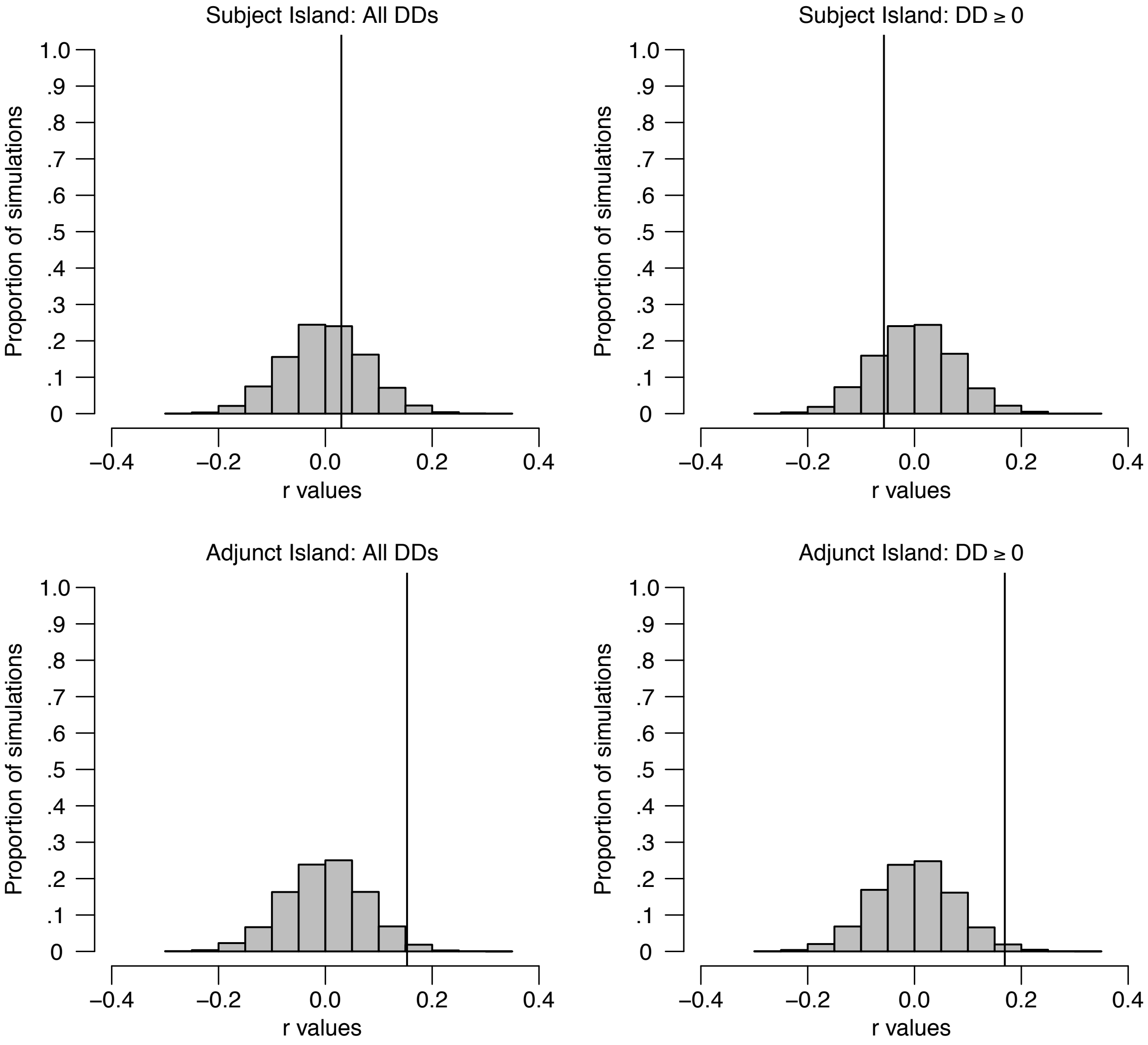
5. We then chose one of the DD scores for each participant, and input those DD scores into a linear regression. Again, this is just like our original analysis, as it is a linear regression based on 173 pairs of DD scores and WM scores (161 DD scores in the case of *n*-back). The difference is that the 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 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 10: Results of the resampling simulations for correlation (Pearson's r) between DD scores and recall scores in Experiment 2 (10,000 simulations). The solid line represents the r-value of the actual sample.





As Figure 10 makes clear, the most frequent results of these resampling simulations result in r values at or very near to 0, suggesting that for the majority of possible samples, there is no correlation whatsoever between DD scores and recall scores. Furthermore, the largest possible negative r values (i.e. the largest possible r values in the direction predicted by resource-limitation theories) for any of the island types is approximately -.35, and occurred in less than 1% of the simulations (and only in the Subject islands after eliminating participants with negative DD scores). This suggests that the best possible outcome given our results is for memory resources to explain 12.25% (the R^2 for an r-value of -.35) of the variation in acceptability judgments of island effects, and that would only occur in less than 1% of experiments. From our perspective, this is a much weaker relationship than would be predicted by resource-limitation theories.

7. GENERAL DISCUSSION. In the previous sections, we presented two experiments designed to investigate one of the primary predictions of the resource-limitation theory of island effects (Kluender and Kutas 1993, Kluender 1998, 2004, Hofmeister and Sag 2010). This theory predicts that the strength of island effects should correlate with individual working memory capacity. We operationalized the notion of an ‘island effect’ as the size of the super-additive interaction in acceptability judgments in a factorial definition of island effects. The results of those experiments, as revealed by linear regressions and resampling simulations, suggest no evidence of a relationship between working memory capacity and the strength of the superadditive interaction. Clearly, these results run contrary to the prediction of the resource-limitation theory. However, these results are negative in nature: there is no evidence of a relationship between working memory and island effects. The question then is how strongly this result argues against the viability of the resource-limitation theory as an account of island effects. In this section, we examine two broad types of concerns that might arise in evaluating the consequences of these results for the resource-limitation theory, in an attempt to evaluate the strength of the inferences that can be made.

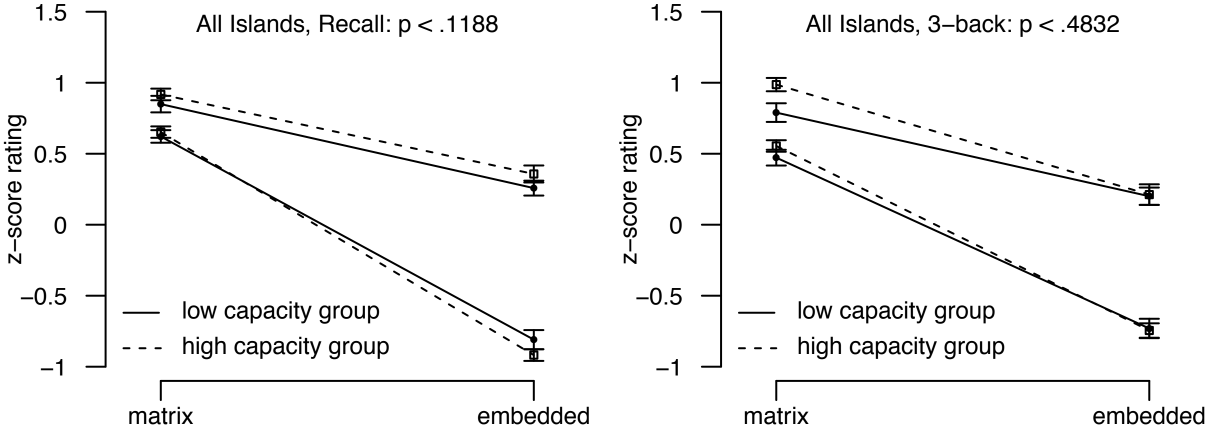
7.1. POTENTIAL CONCERNS ABOUT THE STATISTICAL ANALYSES. One possibility is that there is in fact a significant relationship between working memory capacity and island effects in our data, but that we did not perform the correct statistical analyses to reveal it. We have attempted to guard against this concern by performing every potentially relevant analysis that we could imagine: regressions on z-score transformed responses, regressions on raw responses, regressions on all of the DD scores, regressions only on DD scores that were greater than or equal to zero, regressions with single capacity scores, regressions with multiple capacity scores, and finally resampling simulations. Though we believe that these analyses exhaust the most likely possible avenues for detecting a significant relationship in the data, we concede that there may be other relevant analyses that we have not considered. As such, we have made the complete data set available for download on the first author’s website.Another concern that may arise revolves around the unique logical structure of null hypothesis significance testing (NHST). As is well known, NHST cannot be used to prove the null hypothesis: whereas small p-values can be interpreted as evidence against the null hypothesis, large p-values cannot be interpreted as evidence for the null hypothesis. The reason for this is straightforward: NHST assumes the null hypothesis is true when calculating the p-values. In other words, NHST answers the question If the null hypothesis were true (i.e. there is no difference between conditions), how likely would these results be? The p-value is a measure of the likelihood of the results in a world in which the null hypothesis is true. Because the null hypothesis is assumed during the calculation of p-values, p-values can’t be used to establish the null hypothesis. Given that the results of our experiments are null, one could worry that we face the same problem of “proving the null”. In fact, such a concern would be misplaced. We did not rely upon any NHST tests in the process of evaluating our results. The least-squares procedure used to derive the lines-of-best fit and the calculation of R^2 values are both descriptive statistics similar to calculating the mean of a set of data points. Although NHST can be applied to the values of these procedures (similar to the way that NHST can be applied to means), it is not necessary to apply NHST to derive information from them (similar to the way that NHST is unnecessary to interpret a mean).

7.2. POTENTIAL CONCERNS ABOUT WORKING MEMORY. Another potential concern could be that we only tested two working memory capacity measures, so it is logically possible that a different capacity measure could be found that does indeed correlate with the acceptability of island effects. As discussed in Section 3, we attempted to guard against this possibility by selecting our tasks in a way that maximizes coverage of the working memory system while minimizing potential confounds with the acceptability judgment task. The fact is that many working memory measures share common variance (Conway et al. 2005). This concern would therefore be corroborated if one could find a working memory measure that is simultaneously highly correlated with island effects, but not correlated with serial recall, and also not correlated with n-back. Given that research that 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.

7.3. POTENTIAL CONCERNS ABOUT THE THEORY OF WORKING MEMORY ASSUMED BY RESOURCE-LIMITATION THEORIES. Another possible concern could revolve around the theory of working memory endorsed by resource-limitation theories of island effects. Some theories of filler-gap dependency completion reject the assumption of the resource-limitation theory that maintenance of the wh-word is required, or that it is a key predictor of processing difficulty (Fodor 1978, McElree et al. 2003, Vasishth and Lewis 2006, Wagers 2012). This could suggest that the best candidate processes for a resource-limitation 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 memory capacity measures used in this study tracked the former rather than the latter capacity. However, once again the established facts of memory capacity measures make this unlikely. Even simple measures of individual differences in working memory, like serial recall, likely tap into individual differences related to both maintenance and retrieval (such as interference robustness (Unsworth & Engle, 2007)). Therefore resource-limitation theories may shift their emphasis on component memory processes, but the tasks necessary to measure resource capacity for those processes are unlikely to differ sufficiently from the ones employed in the present study to uncover a relation between resource capacity and island effects.

7.4. POTENTIAL CONCERNS ABOUT THE RELATIONSHIP BETWEEN MEMORY CAPACITY MEASURES AND ACCEPTABILITY JUDGMENTS. Though the strength of the island effect did not covary with recall or *n*-back scores, other patterns in the data militate against the conclusion that there are simply no correlations between acceptability judgments and individual difference scores. To illustrate this point, we provide one final analysis that considered only the individuals with the highest and lowest recall and *n*-back scores. We isolated the 22 participants with the highest serial recall scores and the 22 participants with the lowest serial recall scores, assigning them to the high and low recall groups respectively. This corresponded to (approximately) 25% of the participants in Experiment 2. We then repeated the process for the 3-back task, leading to a set of participants corresponding to approximately 27% of the participants. We then performed statistical analyses on each memory measure using the groups (high and low) as a critical factor. In order to increase the likelihood of finding a robust pattern of results, we collapsed the four island effects together into a single design. In other words, we treated every NON-ISLAND | MATRIX condition as identical regardless of island type, we treated every NON-ISLAND | EMBEDDED condition as identical regardless of island type, and so on for the other two condition types. The mean ratings and standard errors of this collapsing procedure are presented in Figure 11 (left panel for serial recall groups, right panel for 3-back groups). These graphs are very similar to the 2×2 interaction graphs presented in previous sections; however, because there is an extra two-level factor (GROUP: high and low), there are now eight points arranged as if two 2x2 interaction graphs are superimposed.

Figure 11: Experiment 2, interaction plots based on two memory groups (high performance and low performance) collapsed over all four island type (n=22 per group per memory measure). The *p*-values at the top of each graph represent the three-way interaction term (STRUCTURE x GAP-POSITION x GROUP).



As a first analysis for each memory type, we constructed a linear mixed-effects model using STRUCTURE (non-island and island), GAP-POSITION (matrix and embedded), and GROUP (high or low) as fixed factors, and participants and items as random effects (equivalent to a three-way, or 2x2x2, ANOVA). The three-factor interaction term (STRUCTURE X GAP-POSITION X GROUP) indicates whether the size of the island effect (i.e. the two-way interaction STRUCTURE X GAP-POSITION) for the high group is significantly different than the size of the island effect for the low group. The *p*-values for the three-factor interaction terms are reported in the top-middle of the graphs in Figure 11: there were no significant interactions of STRUCTURE X GAP-POSITION X GROUP for either memory task, suggesting that the size of the island effects for the high and low groups are not significantly different.

However, based on the visible differences between high and low groups with respect to NON-ISLAND | SHORT conditions in the 3-back task, we decided to run an additional analysis to test for differences between the two groups with respect to the size of the dependency length effect (which could potentially be sensitive to memory capacity depending on one's assumptions about the relationship between memory capacity and acceptability judgments). We constructed a linear mixed-effects model using GAP-POSITION (MATRIX and EMBEDDED) and GROUP (HIGH or LOW) as fixed factors, and participants and items as random effects (equivalent to a two-way, or 2x2, ANOVA). The interaction term (GAP-POSITION X GROUP) indicates whether the size of the dependency length effect (i.e. the pairwise difference between non-island | matrix and non-island | embedded conditions) for the high group is significantly different than the size of the dependency length effect for the low group. The interaction was non-significant for recall scores (*p*=.706), but significant for 3-back scores (p=.036).

As is clear from Figure 11 and the results of the second set of statistical tests, the high-performers in the 3-back task have a significantly larger dependency length effect. In some ways this is counter-intuitive, as one may have predicted that participants with higher memory capacity would be affected less by the dependency length manipulation. However, it is clear in Figure 11 that the difference between the two groups is restricted to the NON-ISLAND | MATRIX condition; this may suggest that the higher performing participants use the response scale slightly differently than low performing participants, perhaps leading to a stronger classification of this condition as highly acceptable. At the very least these results suggest that there is at least one significant relationship between an individual differences measure (*n*-back) and acceptability judgments (the dependency length effect).

7.5. CONCLUSIONS REGARDING RESOURCE-LIMITATION THEORIES. At this point, it seems to us that the only viable avenue for resource-limitation is theories is to assume that the experiments presented here fail on two fronts simultaneously: (i) the current experiments tested the wrong version of the resource-limitation theory (i.e. the correct theory is one that does not rely on any of the theoretical constructs tested by the serial recall and *n*-back tasks), and (ii) the experiments employed the wrong working memory tasks (i.e. the correct theory requires a task that is uncorrelated with both serial recall and *n*-back tasks). In other words, the real explanation for island effects is a currently unformulated theory of working memory that must be tested using a currently unknown measure of working memory capacity. We admit that this is certainly a logical possibility, but considerable caution should be observed in pursuing this explanation. In addition to discounting the results presented here, such an explanation would also undermine the reductionist nature of the resource-limitation theories. The attraction of reductionist theories of islands is that they eliminate the need for cognitive constructs that are postulated solely to explain the existence of island effects. If the "processing resources" necessary to explain island effects turn out not to be independently motivated, then there is little advantage to postulating them over a grammatical constraint.

8. CONCLUSION. In this article, we have argued that the key difference between the resource-limitation reductionist theory and grammatical theories lies in the how they account for the statistical interaction in acceptability scores that arises when dependency length and structure are independently manipulated: the resource-limitation 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 a consequence of a constraint that impacts judgments on only one of the four conditions in the design. Therefore the resource-limitation 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 studies 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 resampling 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 resource-limitation 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 grammatical constraints in the minds of contemporary speakers. In other words, the synchronic cognitive commitments of the grounded theory with respect to island constraints are identical to grammatical theories. However, the results are incompatible with a resource-limitation reductionist theory.

We believe that the results of the experiments presented in this article provide strong support for grammatical theories of island effects because 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 grammatical or grounded theories.

REFERENCES

ABRUSÁN, MÁRTA. 2011. Presuppositional and negative islands: A semantic account. *Natural Language Semantics* 19.257–321.

BAAYEN, R. HARALD. 2007. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

BAAYEN, R. HARALD; DOUGLAS J. DAVIDSON; and DOUGLAS M. BATES. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412.

BARD, ELLEN GURMAN; DAN ROBERTSON; and ANTONELLA SORACE. 1996. Magnitude estimation of linguistic acceptability. *Language* 72.32–68.

BERWICK, ROBERT C., and AMY S. WEINBERG. 1984. *The grammatical basis of linguistic performance*. Cambridge, MA: The MIT Press.

CAPLAN, DAVID, and GLORIA S. WATERS. 1999. Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences* 22.77–94.

CHOMSKY, NOAM. 1973. Conditions on transformations. *A Festschrift for Morris Halle*, ed. by Stephen Anderson and and Paul Kiparsky, 232–286. New York: Holt, Rinehart, and Winston.

CHOMSKY, NOAM. 1986. *Barriers*. Cambridge, MA: The MIT Press.

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

CONWAY, ANDREW. R. A., MICHAEL J. KANE, MICHAEL F. BUNTING, D. ZACH HAMBRICK, OLIVER WILHELM, and RANDALL W. ENGLE. 2005. Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review* 12.769–786.

DANEMAN, MEREDYTH. and PATRICIA A. CARPENTER. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior* 19.450–466.

DEANE, PAUL. 1991. Limits to attention: a cognitive theory of island phenomena. *Cognitive Linguistics* 2.1–63.

EDGINGTON, EUGENE, and PATRICK ONGHENA. 2007. *Randomization tests* (4th ed.). Boca Raton, FL: Chapman and Hall/CRC.

ENGDAHL, ELISABET. 1983. Parasitic Gaps. *Linguistic Inquiry* 6.5–34.

ERTESCHIK-SHIR, NOMI. 1973. *On the nature of island constriaints*. Cambridge, MA: MIT dissertation.

FEATHERSTON, SAM. 2005a. Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115.1525–1550.

FEATHERSTON, SAM. 2005b. Universals and grammaticality: Wh-constraints in German and English. *Linguistics* 43.667–711.

FEDORENKO, EVELINA; EDWARD GIBSON; and DOUGLAS ROHDE. 2006. The nature of working memory capacity in sentence comprehension: Evidence against domain specific resources. *Journal of Memory and Language* 54.541–553.

FEDORENKO, EVELINA; EDWARD GIBSON; and DOUGLAS ROHDE. 2007. The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language* 56.246–269.

FODOR, JANET D. 1978. Parsing strategies and constraints on transformations. *Linguistic Inquiry* 9.427–473.

FODOR, JANET. D. 1983. Phrase structure parsing and the island constraints. *Linguistics and Philosophy* 6.163-223.

FORSTER, KENNETH I. AND JONATHAN C. FORSTER. 2003. DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods* 35.116–124.

GIVÓN, TALMY. 1979. *On Understanding Grammar.* New York: Academic Press.

GOLDBERG, ADELE. 2007. Constructions at work. Oxford University Press.

HAWKINS, JOHN A. 1999. Processing complexity and filler-gap dependencies across grammars. Language 75.244–285.

HOFMEISTER, PHILIP, AND IVAN A. SAG. 2010. Cognitive constraints and island effects. *Language* 86.366–415.

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

JAEGGI, SUSANNE M.; MARTIN BUSCHKUEHL; JOHN JONIDES; and WALTER PERRIG. 2008. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences* 105.6829–6833.

JOLLIFFE, IAN T. 2002. Principal Component Analysis. New York: Springer-Verlag.

JUST, MARCEL A., and PATRICIA A. CARPENTER. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 98.122–149.

KANE, MICHAEL J., and RANDALL 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, MICHAEL J., ANDREW R. A. CONWAY, TIMOTHY K. MIURA, and GREGORY 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* 33.615–622.

KELLER, FRANK. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Edinburgh: University of Edinburgh dissertation.

KELLER, FRANK. 2003. A Psychophysical Law for Linguistic Judgments. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, ed. by Richard Alterman and David Kirsh, 652–657. Mahwah: New Jersey: Lawrence Erlbaum.

KING, JONATHAN, and MARCEL A. JUST. 1991. Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language* 30.580–602.

KIRCHNER, WAYNE K. 1958. Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology* 55.352–358.

KLUENDER, ROBERT, and MARTA KUTAS. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8.573–633.

KLUENDER, ROBERT. 1998. On the distinction between strong and weak islands: A processing perspective. *Syntax and Semantics 29: The Limits of Syntax*, ed. by Peter Culicover and Louise McNally, 241–279. New York: Academic Press.

KLUENDER, ROBERT. 2004. Are Subject islands subject to a processing account? *Proceedings of the West Coast Conference on Formal Linguistics 23*, ed. by Vineeta Chand, Ann Kelleher, Angelo Rodriguez, and Benjamin Schmeiser, 101–125. Somerville, MA: Cascadilla Press.

KUNO, SUSUMU. 1976. Subject, Theme, and the Speaker's Empathy - A Reexamination of Relativization Phenomena. *Subject and Topic*, ed. by Charles N. Li, 417–44. New York: Academic Press.

KUNO, SUSUMU. 1987. *Functional Syntax: Anaphora, Discourse and Empathy*. Chicago: University of Chicago Press.

KUNO, SUSUMU, and KEN-ICHI TAKAMI. 1993. *Grammar and Discourse Principles: Functional Syntax and GB Theory.* Chicago: University of Chicago Press.

LASNIK, HOWARD, AND MAMURO SAITO. 1984. On the nature of proper government. *Linguistic Inquiry* 15.235–289.

MACDONALD, MARYELLEN. C.; MARCEL A. JUST; and PATRICIA A. CARPENTER. 1992. Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology* 24.56–98.

MACMILLAN, NEAL A., and C. DOUGLAS CREELMAN. 2004. *Detection Theory: A User's Guide*. Mahwah, N.J.: Lawrence Erlbaum Associates.

MAXWELL, SCOTT E., and HAROLD D. DELANEY. 2003. *Designing Experiments and Analyzing Data: A model comparison perspective*. Mahwah, N.J.: Lawrence Erlbaum Associates.

MCELREE, BRIAN; STEPHANI FORAKER; and LISBETH DYER. 2003. Memory structures that subserve sentence comprehension. *Journal of Memory and Language* 48.67–91.

PABLOS, LETICIA. 2006. *Pre-verbal Structure Building in Romance Languages and Basque*. College Park, MD: University of Maryland Dissertation.

PEARSON, KARL. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 6.559–572.

PEARL, LISA and JON SPROUSE. 2012. Computational Models of Acquisition for Islands. *Experimental Syntax and Island Effects*, ed. by Jon Sprouse and Norbert Hornstein. Cambridge University Press.

PHILLIPS, COLIN. 2006. The real-time status of island constraints. *Language* 82.795–823.

PRITCHETT, BRADLEY. 1991. Subjacency in a Principle-Based Parser. *Principle-Based Parsing: Computation and Psycholinguistics*, ed. by Robert C. Berwick, Steven P. Abney, and Carol Tenney, 301–345. Dordrecht: Kluwer.

RIZZI, LUIGI. 1982. Violations of the wh-island constraint and the subjacency condition. *Issues in Italian Syntax,* ed. by Luigi Rizzi, 49–76. Dordrecht, NL: Foris.

ROBERTS, ROSE. and EDWARD GIBSON. 2002. Individual differences in working memory. *Journal of Psycholinguistic Research* 31.573–598.

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, JOHN ROBERT. 1967. *Constraints on variables in syntax*. Cambridge, MA: MIT dissertation.

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

Sprouse, Jon. 2008. The differential sensitivity of acceptability to processing effects. *Linguistic Inquiry* 39.686–694.

Sprouse, Jon. 2009. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*. 40.329–341.

Sprouse, Jon; Shin Fukuda; Hajime Ono; and Robert Kluender. 2011. Grammatical operations, parsing processes, and the nature of wh-dependencies in English and Japanese. *Syntax* 14.179–203.

Sprouse, Jon. 2011. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87.274–288.

Stevens, Stanley Smith. 1956. The direct estimation of sensory magnitudes: loudness. *The American journal of psychology* 69.1–25.

Szabolcsi, Anna, and Marcel den Dikken. 2006. Strong and weak islands. *The Blackwell Companion to Syntax*, ed. by Martin Everaert and Henk van Riemsdijk, 479–532. Somerset, NJ: Wiley-Blackwell.

Szabolcsi Anna, and Frans Zwarts. 1993. Weak islands and an algebraic semantics of scope taking. *Natural Language Semantics* 1.235–284.

Torrego, Ester. 1984. On inversion in Spanish and some of its effects. *Linguistic Inquiry* 15.103-129.

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

Ueno, Mieko, and Robert Kluender. 2009. On the processing of Japanese wh-Questions: An ERP study. *Brain Research* 1290.63–90.

Unsworth, Nash, and Randall W. Engle. 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, Sandra H.; Thomas C. Gunter; Herbert Schriefers; and Angela D. Friederici. 2001. Syntactic parsing and working memory: The effects of syntactic complexity, reading span, and concurrent load. *Language and Cognitive Processes* 16.65–103.

Weskott, Thomas, and Gisbert Fanselow. 2011. On the Informativity of Different Measures of Linguistic Acceptability. *Language* 87.249–273.

APPENDIX OF ADDITIONAL 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 ≥ 0	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 ≥ 0	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 ≥ 0	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 ≥ 0	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	.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 ≥ 0	Intercept	0.91	18.26	.001	-.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 ≥ 0	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	-.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 ≥ 0	Intercept	1.44	25.53	.001	-.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	.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 ≥ 0	Intercept	0.95	19.45	.001	.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		