Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

2 Deriving competing predictions from grammatical approaches and reductionist approaches to island effects

Jon Sprouse, Matthew W. Wagers, and Colin Phillips

1 What is the relationship between grammatical theories and parsing theories?

Marr (1982) famously proposed that our theories of information-processing devices can be usefully stated at multiple levels: the computational level, the representational-algorithmic level, and the implementational level. Marr described the computational level as an answer to the question "What problem must this device solve?" He argued that the computational level would specify the properties of the problem that must be solved by the device and the computations that the device must perform in service of that goal, in a way that abstracts away from the exigencies of actually solving the problem in practice. Marr used a cash register as an example: the computational-level description of a cash register comprises the theory of addition, including properties such as commutativity and associativity. However, at the computational level there is no statement of the procedure the device follows or the series of states it occupies to carry out addition. A theory at that level of description is a representational-algorithmic theory. For a cash register this could be the addition algorithm that we all learn in school, implemented in base 10: start from the right, and "carry over the ones"; or it could be implemented in base 2, which a digital device would use. Finally, Marr described the implementational level as a theory of how the operations of the algorithmic level are implemented in the hardware of the device. For a cash register, there are several hardware options that can implement this level, from the spinning drums in mechanical cash registers to the electronic processors in computers.

Extending the Marr framework to sentence-level language phenomena is relatively straightforward, at least in theory. Grammatical theories tend to be computational-level descriptions, as they describe the properties of the final grammatical structures that must be built, as well as the properties of the structure-building operations that are required to build them, but abstract away from the requirements of real-time sentence processing. Parsing theories tend to be algorithmic-level theories, as they describe the specific parsing

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

operations that must be deployed during real-time sentence processing, including the strategies that dictate the deployment of those operations, and the ways in which parsing resources constrain the operation of the parser. Finally, at the implementational level there are currently theories of the macroscopic organization of the brain areas that subserve language processing and the ways in which they interact as well as computer models of neural-like computation. Viewed in this way, it is clear that each of the levels is in fact a description of the same object (the human sentence-processing system), just viewed from slightly different perspectives. Each level brings certain properties of the system into focus, while abstracting away from other properties of the system.

Though the debate between grammatical approaches to island effects and "processing" approaches to island effects appears to set up a dichotomy between the grammar and the parser, the discussion of Marr's levels above illustrates that no such dichotomy need exist in the actual human sentence-processing system. The first step in addressing the question of the source of island effects is to be as explicit as possible about which components of the sentence-processing system the grammatical theory is intended to describe, and which components of the sentence-processing system the parsing theory is intended to describe. In this way, we can isolate the real differences between the two levels of description, and determine exactly what is meant by grammatical approaches to island effects and "processing" approaches to island effects. We will attempt such an elaboration in section 2.

As with all theoretical questions, researchers may disagree about the precise relationship between grammatical theories and parsing theories. We assume that a complete parsing theory will at least include a set of (incremental) structure-building operations, a set of control mechanisms that determine which structure-building operations should be deployed in each environment, and a set of resources (such as working memory) that are recruited to deploy the operations (cf. Lewis 2000). We assume that grammatical theories are a redescription of the structure-building component of the theory that abstracts away from parsing strategies and resource requirements, and also abstracts away from the exigencies of parsing speech (i.e., left-to-right directionality, error detection, etc.). The in-principle existence of an indirect relationship between grammatical theories and parsing theories complicates how specific operations of the parser can be linked to properties of the grammatical theory. In particular, because parser processes and grammatical computations need not map in one-to-one fashion, then a body of observations about language comprehension performance are likely compatible with more than one grammatical theory.

2 The logic of the reductionist approach

The terms 'grammatical explanation' and 'processing explanation' are used so frequently in the islands literature that it seems as though it should be

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

obvious what exactly they refer to. The prevalent intuitive definition appears to be something like this: a grammatical explanation assumes the existence of a grammatical constraint (often syntactic, but not necessarily so) to explain island effects, whereas a processing explanation assumes a constraint of the parsing system. Given the discussion of the relationship between grammatical theories and parsing theories above, it should now be clear that this terminology admits a certain amount of ambiguity. The many-to-one relationship between parsing theories and grammatical theories means that in order for there to be a useful distinction between grammatical explanations and processing explanations, the processing explanations must be restricted to the components of the sentence-processing system that are not also described by grammatical theories. One way to think about this is as a distinction between (i) structure-building operations that are illegitimate, and (ii) structure-building operations that are possible, but aren't carried out in specific circumstances due to constraints on the resources available to the parsing system. Explanations of type (i) refer to constraints that occur in the processing system that mirror constraints on the computations of the grammatical theory, whereas explanations of type (ii) refer to constraints that occur in the processing system as a consequence of its existence in time and space, where resources are finite and the input representations may be noisy.

Because of the terminological ambiguity discussed above, Phillips (2013) and Sprouse *et al*. (2012) suggest that processing explanations should instead be called *reductionist explanations*. The term *reductionism* better captures the underlying logic of the processing-based approach, which in turn helps to clarify how the two approaches can be teased apart. Crucially, reductionist explanations seek to reduce island effects to one or more components of the sentence-processing system that are motivated by language-independent perceptual or cognitive properties. A reductionist explanation thus decreases the number of formal grammatical constraints that must be postulated to capture the phenomena of a given language. This logic has two important consequences. First, reductionist explanations must explicitly specify which mechanisms give rise to the island effect. As mentioned above, these mechanisms must not be structure-building operations or abstract constraints on structure-building operations, otherwise they are equivalent to an explanation in grammatical terms. Second, the mechanisms of the reductionist explanation must be independently motivated; in other words, the mechanisms should be necessary to explain phenomena other than island effects. If the mechanisms are not independently necessary, then the reductionist explanation is not truly reductionist.

Recasting the debate in terms of grammatical explanations and reductionist explanations also allows us to articulate additional distinctions among theories of island effects. For example, there is a set of theories that combine aspects of both grammatical and reductionist accounts, which Sprouse *et al*. (2012) call *grounded theories*. Grounded theories share with grammatical theories the assumption that island effects are caused by grammatical constraints within a

UNCORRECTED PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

speaker's mind; however, they also share with reductionist theories the assumption that island effects arise because these particular structures, if they were generated, would be difficult to parse. The guiding intuition is that the inherent difficulty of these structures has led to the grammaticization of a set of island constraints over the course of the history of the language because such constraints bestow an adaptive advantage to the synchronic speaker. Classic examples of grounded theories are Berwick and Weinberg (1984) and Hawkins (1999). For the purposes of this chapter, we will focus on pure grammatical and reductionist theories, leaving non-formal, non-reductionist theories and grounded theories for future research.

3 The factorial definition of island effects

Armed with the distinction between grammatical and reductionist explanations, we are now in a position to define island effects in a way that will allow us to tease apart the two types of theories. Perhaps the most common definition of island effects in the syntactic literature is the *absolute* definition. The absolute definition involves only a single sentence, the island-violating sentence:

(1) *What do you wonder whether John bought __?

Under the absolute definition, an island effect is simply a severe level of unacceptability (i.e., below some threshold set by the researcher) associated with long-distance dependencies out of island structures. A second popular definition of island effects is the *relative* definition: by comparing the island-violating sentence with an appropriately matched grammatical sentence, the island effect can be defined as a relative difference in acceptability between the two sentences. For example, a common control condition for *whether* islands is a long-distance dependency out of a CP headed by *that*:

(2) a. What do you think [that John bought __]?
b. *What do you wonder [whether John bought __]?

It is easy to see a progression between these two definitions. Whereas the absolute definition establishes that the island violation sentence is indeed unacceptable, it is not clear what the source of the unacceptability is. For example, it is possible that long-distance dependencies across clause boundaries are simply unacceptable in English. This is not true, but the absolute definition does not give us this information. The relative definition corrects this flaw by illustrating that a long-distance dependency that crosses a CP headed by *that* is indeed acceptable. However, in the process, a second confound becomes apparent: it is possible that the unacceptability of the island-violating sentence is driven by the mere presence of *whether*, rather than by the location of the gap inside of the island structure. To control for this possibility, we can add to the

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

paradigm a third condition that contains a CP introduced by *whether* without a long-distance dependency out of the embedded clause:

(3) a. What do you think [that John bought __]?
b. Who __ wonders [whether John bought a car]?
c. *What do you wonder [whether John bought __]?

The triplet in (3) is sufficient to establish (logically) that the unacceptability in (3c) is unique to the combination of two properties: an embedded clause introduced by *whether* plus a long-distance dependency out of the embedded clause. The acceptability of (3a) and (3b) jointly demonstrate that neither property alone is sufficient to cause unacceptability.

At this point, it is clear that we are manipulating two factors: the structure of the embedded clause (STRUCTURE), and the position of the gap (GAP). Each factor has two levels: (ISLAND/NON-ISLAND) and (MATRIX/EMBEDDED) respectively. By crossing both factors, we obtain the set of sentences below:

(4) 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

In this design, condition (4a) serves as a baseline as it is a combination of the "unmarked" levels of each factor. Condition (4b) manipulates the location of the dependency such that difference in acceptability between (4a) and (4b) isolates the effect of long-distance wh-movement. The difference between (4a) and (4c) isolates the effect of *whether* clauses. Finally, the acceptability difference between (4a) and (4d) represents the combination of the two factors, as (4a) is the baseline sentence containing neither an island domain nor a non-local extraction while (4d) is the sentence containing both. Condition (4a) is thus crucial, because it serves as a baseline that allows us to isolate the effect of each of the factors in this design. Reductionist theories posit that island effects are a combination of independently motivated effects and therefore it is necessary to factor out these individual effects to assess the empirical plausibility of specific reductionist accounts.

4 The simplest reductionist account: a linearly additive effect

As a first pass, we can construct the simplest possible reductionist account and see how we can use the factorial definition to assess its empirical plausibility. We will consider the *whether* island we have been using as an example. The first assumption is that processing costs such as the taxation of working memory are directly reflected in acceptability judgments. In other words, processing costs lead to lower acceptability judgments. The second assumption is that there are

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

cognitive resource costs associated with parsing long-distance dependencies, such as might derive from strains on working memory or attention. This is a relatively common assumption in the sentence-processing literature, although the details of the mechanisms vary considerably from theory to theory (see Wagers, this volume, and references therein). The third assumption is that there is a processing cost associated with the construction of embedded *whether* clauses. Though this putative cost is not widely discussed or investigated in the sentence-processing literature, it is straightforward to imagine that the complex semantics associated with embedded questions could entail some sort of processing cost at the semantic or discourse level (Kluender and Kutas 1993b). With these three assumptions in hand, we now have the ingredients for a simple reductionist theory: we have two independently motivated processing costs (the cost of long-distance dependencies and the cost of *whether* clauses), and we have a linking hypothesis between processing costs and acceptability judgments. The theory takes the following form: Each individual processing cost is small enough that sentences containing 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 (no asterisk) from unacceptable sentences (asterisk).

The simple reductionist theory outlined above makes strong predictions regarding the numerical ratings given to each of the four conditions in the (fully crossed) factorial design. In particular, this theory predicts that the relationship between the two processing costs (long-distance dependencies and the construction of *whether* clauses) should be linearly additive: the cost of processing long-distance dependences [(4a)-(4b)] plus the cost of processing *whether* clauses [(4a)-(4c)] should equal the cost of performing both together [(4a)-(4d)]. In formula form: [(4a)-(4b)] + [(4a)-(4c)] = [(4a)-(4d)]. This prediction can be graphically represented using an interaction plot (Figure 2.1).

Crucially, a linearly additive relationship within a 2×2 design results in parallel lines. Given the arrangement used in Figure 2.1, the separation between the two lines represents the main effect of *whether* clauses, and the slope of the lines represents the main effect of long-distance dependencies. The rating of the island-violating sentence (condition (4d), which is in the bottom right quadrant of Figure 2.1) is simply the sum of these two values. In this way, there is no need to invoke an additional grammatical 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 factorial definition in (4) has been used to test several island types in English by Sprouse (2007a), Sprouse *et al*. (2011), and Sprouse *et al*. (2012). Figure 2.2 reports the results for *Whether*, CNPC, Adjunct, and Subject islands from Sprouse *et al*. (2012).

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

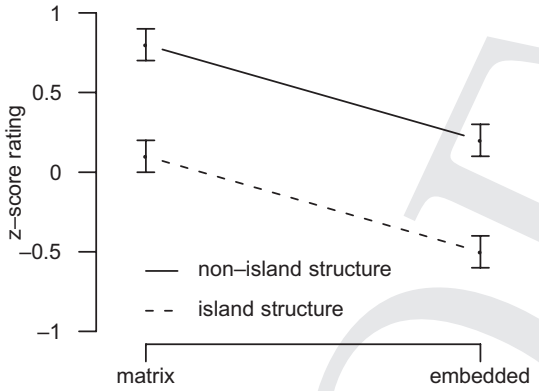

Figure 2.1 A graphical example of a linearly additive effect with a 2×2 design

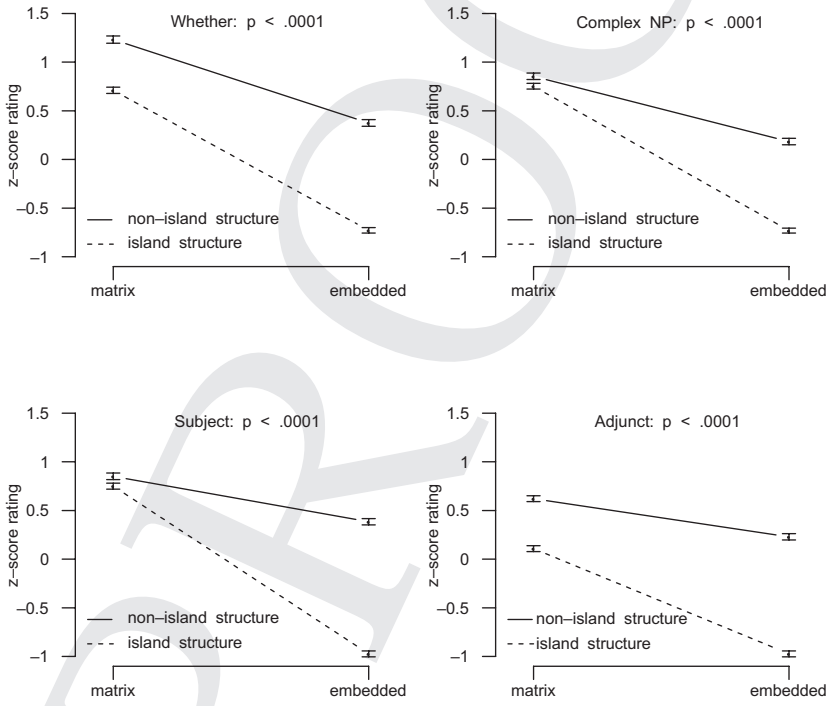

Figure 2.2 Results of magnitude estimation experiments for four island types from Sprouse *et al.* (2012). The p-value of the interaction term is at the top of each graph. N = 173.

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

It is clear from the non-parallelism of the pairs of lines in Figure 2.2 that the combined effect of the two costs in each plot is greater than the (linear) sum of the individual costs; in other words: [(4a)-(4b)] + [(4a)-(4c)] < [(4a)-(4d)]. In other words, the island effect is *superadditive*, as the whole is greater than the sum of the parts. Statistically, this superadditive effect tends to manifest as an interaction, reflecting the fact that the response to each level of each factor is dependent upon the level of the other factor. Interactions can be quickly identified in figures by the non-parallel lines. Crucially for our purposes, the superadditive effect that arises with island effects using the factorial design in (4) suggests that the simple reductionist theory sketched above is an empirically inadequate description of the actual facts of English. Therefore an additional component must be added to the explanation to account for the superadditive effect. This extra component can be a syntactic or semantic constraint that targets condition (4d) and lowers its acceptability below the linear sum predicted by the two factors. Alternatively, this extra component could be a linking hypothesis that causes the two component processing costs to interact in a way that leads to the superadditive effect (which we would call a "processing" explanation). In other words, there's no question at this point that the factors combine superadditively. The question is what is responsible for this effect.

5 Superadditive effects and working memory: an elaborated reductionist account

It should be clear at this point that the space of possible answers to our new driving question (*What causes the superadditive effect in Figure 2.2?*) is very large. One can propose any number of (grammatical) constraints that target condition (4d), and indeed, many have been proposed (Chomsky 1973, Huang 1982a, Lasnik and Saito 1984, Chomsky 1986, Rizzi 1990, Szabolcsi and Zwarts 1993, Tsai 1994, Reinhart 1997, Hagstrom 1998, Chomsky 2000, Truswell 2007, and many others). Similarly, one can propose any number of reasons that the processing costs of building long-distance dependencies might interact with the processing of island structures; however, unlike grammatical explanations, only one such theory has been proposed: the *resource capacity* theory of Kluender and Kutas (1993b). For the remainder of this chapter we will focus on the resource capacity theory as a case study in assessing the empirical plausibility of reductionist explanations of island effects. The resource capacity theory of Kluender and Kutas (1993b) is the only reductionist account that contains a mechanism to explain the superadditive effect of islands. Furthermore, while it is true that there could be several different explanations for the interaction between the processing of long-distance dependencies and island structures (of which the resource capacity theory is just one), it is simply not the case that there are several different ways to factor island effects. The STRUCTURE and

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

GAP factors that are crucial to the resource capacity theory are likely to be the only factorization of island effects possible in English; as such, the general form of the resource capacity theory should hold for any future reductionist proposals.

The resource capacity theory of Kluender and Kutas (1993b) works as follows (see also Kluender 1998, 2004, and Hofmeister and Sag 2010 for elaborations). First, it is assumed that every human comes to the sentence-processing task with a limited amount of working memory with which to manage sentence processing (see Wagers, this volume, for discussion). Second, each of the component processes that we have been discussing (the processing of long-distance dependencies and the processing of island structures) is assumed to be associated with a working memory resource requirement. Third, although the working memory requirements of each process are assumed to be within the limits available to the speaker, it is assumed that the sum of the requirements *exceeds* these limits. By consuming all of the comprehender's resources, the simultaneous deployment of the two sets of processes will lead either to outright failure or intolerably slow computation. Finally, this theory assumes that the act of exceeding the available capacity will lower acceptability judgments. In this way, the extra unacceptability that characterizes the superadditive effect is simply the penalty for exceeding the amount of resources available to the speaker.

It should be clear at this point that the resource capacity theory of Kluender and Kutas (1993b) is a potentially viable theory of the results in Figure 2.2:[1] the superadditive effect is explained by (i) the link between the two processes (they both use working memory resources) and (ii) the penalty for exceeding the limited pool of working memory resources. Whether the resource capacity theory accurately reflects the mechanisms of processing long-distance dependencies is a different question. One method of assessing any reductionist theory is to investigate each of the mechanisms of the resource capacity theory to determine whether they are independently required by the system. In order for a theory to be truly reductionist, each component must be motivated by phenomena other than island effects; if any of the components are motivated only by the existence of island effects, then it is not a truly reductionist account.

Assumptions of the resource capacity theory (Kluender and Kutas 1993b):
1. There are operations for parsing long-distance dependencies.
2. There are operations for parsing island structures.

[1] The lack of structure effects in CNPC and Subject islands may be problematic for the resource capacity theory insofar as it predicts that each of the independent costs should yield (not insubstantial) differences in the acceptability ratings. This problem is discussed a bit more in section 6. A more complete review of the viability of each of the assumptions of the resource capacity theory is provided in Sprouse *et al*. 2012.

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

3. There is a limited pool of working memory resources.
4. Long-distance dependencies operations require working memory resources.
5. Island structure processes require working memory resources.
6. The sum of the resources required by 4 and 5 is greater than the resources available.
7. There is a parsing algorithm that deploys the operations in 1 and 2 simultaneously.

For reasons of space, we will not go into the independent plausibility of each of the assumptions of the resource capacity theory; we will just say that assumptions 1–4 appear to be relatively uncontroversial (though they are all active areas of research), while assumptions 5–7 are less widely accepted. The interested reader should consult Wagers (this volume) for a detailed discussion of the potential parsing and memory mechanisms involved in the processing of long-distance dependencies with and without island structures.

A second method of assessing any reductionist theory of island effects is to highlight the mechanism that is invoked to explain the superadditive component. Grammatical accounts of islands explain the superadditive effect by targeting the ISLAND | EMBEDDED condition (4d) with a specific rule of grammar. Though this specific rule adds to the number of assumptions of the theory, it has one crucial benefit: it has no effect on the other conditions in the factorial design. In contrast, reductionist explanations by definition must invoke theoretical mechanisms that will affect two or more of the conditions simultaneously. Though the details of how this fact can be leveraged will vary from theory to theory, in all cases it means that the grammatical theory and the reductionist theory should make different predictions regarding manipulations of the theoretical mechanism that explains the superadditive component. This opens the door for testing those competing predictions experimentally.

As a concrete example, Sprouse *et al*. (2012) argue that the primary difference between grammatical explanations and the resource capacity theory lies in the role of working memory capacity: limited working memory capacity is the cause of the superadditive effect under the resource capacity theory, whereas working memory capacity is orthogonal to the superadditive effect under grammatical theories. The mechanism for the superadditive acceptability effect in Kluender and Kutas (1993b) is a type of penalty mechanism: if the resources of the system are exceeded by the demands of the current (simultaneous) operations, the resources allocated to each process are decremented by the amount necessary to bring the system back within the bounds of the resource capacity. In other words, if the capacity is C, the total demand is D, the demand of each process is P, and the number of processes is N, then the resources allocated to each process equal P–(D-C)/N. It is this resource penalty that causes the additional acceptability decrease (the superadditive component). This mechanism suggests a possible prediction for the resource capacity theory: if the

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

capacity (C) is increased, then the penalty on each process will be decreased, thereby decreasing the superadditive component of the acceptability ratings. In other words, the resource capacity theory predicts that working memory capacity should (negatively) correlate with the magnitude of the superadditive effect in the factorial design in (4), whereas grammatical explanations predict that there should be no correlation between working memory capacity and the superadditive effect.[2]

6 Testing the working memory predictions

The first step in testing the competing predictions of the two theories is to derive a measure (or measures) of working memory capacity. There are a variety of ways of measuring an individual's working memory resources (see Roberts and Gibson 2002). The many indexes reflect, in part, the existence of separate cognitive mechanisms underlying processing efficiency and, in part, the fact that there can be multiple ways of operationalizing these mechanisms in experimental tasks. To circumvent the problem of choosing 'the right measure', Sprouse *et al*. (2012) used two short-term memory tasks: the serial recall task and the *n*-back task. These tasks were chosen because the literature on individual differences suggests that a relatively few underlying constructs can account for most of the variance across memory tasks. The serial recall and *n*-back tasks have been shown to be closely related to each of these components, but crucially do not appear to be closely related to each other (Conway *et al*. 2005, Kane *et al*. 2007). Taken together, these two tasks likely cover a large portion of the possible variance in working memory resources, making it unlikely that other memory tasks will lead to different results. Moreover, both measures have been implicated in sentence memory (Roberts and Gibson 2002).

In the serial recall task participants are presented with a series of words one at a time, and when the presentation is complete, they are asked to recall those words in the order that they were presented (see Cowan 2001 and Conway *et al*. 2005 for reviews). Sprouse *et al*. (2012) included features to help eliminate mnemonic strategies, such as asking participants to softly repeat "the" during the trials to inhibit rehearsal, and using the same words (in a different order) in each trial to eliminate unique semantic associations. In the *n*-back task, participants are presented with a series of letters on a computer screen one at a time (rapid serial visual presentation or RSVP), and are asked to

[2] It is logically possible to develop resource capacity theories that differ mechanistically from the Kluender and Kutas (1993b) theory. It is of course an open question whether every possible formulation would make the same prediction regarding the (negative) correlation between working memory capacity and the superadditive component of island ratings. We believe that a large number of formulations would also make this prediction, thus it is a reasonable place to begin the investigation of the resource capacity theory.

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

Table 2.1 *Calculating the DD score with a sample set of mean ratings*

a.	D1 = (long, non-island) – (long, island)		rating (*z*-score units)
	What do you think that John bought ___?		0.5
	What do you wonder whether John bought __?	–	–1.5
			2.0
b.	D2 = (short, non-island) – (short, island)		
	Who __ thinks that John bought a car?		1.5
	Who __ wonders whether John bought a car?	–	0.7
			0.8
c.	DD = D1 – D2 = 2.0 – 0.8 = **1.2**		

press a button if the letter currently on the screen was also presented *n* items previously (Kirchner 1958, Kane and Engle 2002, Jaeggi *et al*. 2008). This means that in order to complete the task successfully, the participant must continuously update the *n* letters that are kept in memory through the entire presentation (in our experiments, 30 letters were presented in sequence during each trial). By increasing the value of *n* (in our experiments, participants completed a 2-back, 3-back, and 4-back task, in that order), the experimenter can increase the difficulty of the task to obtain a working memory capacity measure.

The second step is to derive a measure of the size of the superadditive acceptability effect using the design in (4). Because of the recent interest in the differences (and similarities) among the various acceptability judgment tasks (Bader and Häussler 2010, Sprouse 2011, Weskott and Fanselow 2011, Sprouse and Almeida 2012), Sprouse *et al*. (2012) used two different judgment tasks: 7-point Likert scale and Magnitude Estimation (Stevens 1957, Bard *et al*. 1996). Both of these tasks provide numerical ratings that can be used to determine the superadditive acceptability effect. As one possible analysis, Sprouse *et al*. (2012) used a *differences-in-differences* (DD) score to measure the strength of the superadditive effect for each individual (Maxwell and Delaney 2003). DD scores are calculated for a two-way interaction as follows: First, calculate the difference (D1) between two of the four conditions. To make the DD scores as intuitively meaningful as possible, Sprouse *et al*. (2012) defined D1 as the difference between the embedded, non-island rating and the embedded, island rating. Second, calculate the difference (D2) between the other two conditions. Sprouse *et al*. defined D2 as the difference between the matrix, non-island rating and the matrix, island rating. Finally, calculate the difference between these two difference scores. (See Table 2.1.)

Because DD scores can be calculated for each individual tested (using standard continuous acceptability judgment experiments), DD scores can serve

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

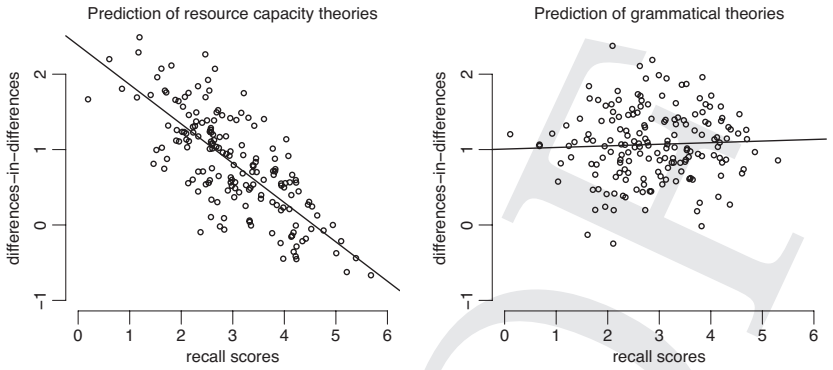

Figure 2.3 Predictions of the capacity-based and grammatical theories

as a composite measure of the strength of the statistical interaction for each individual and intuitively can be thought of as the strength of the island effect for that individual: a positive DD score reflects a superadditive interaction, with larger values representing larger interactions (stronger island effects); a DD score of 0 represents no interaction at all (which is equivalent to no island effect under our specific definition of island effects as a superadditive interaction).

The final step is to couch the predictions of both the resource capacity theory and grammatical theories in terms of the relationship between the strength of the superadditive component (in this case, DD scores) and working memory capacity (in this case, serial recall and *n*-back performance). Sprouse *et al.* argue that the resource capacity theory predicts 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 (i.e., 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 capacity theory predicts that we should see a downward sloping trend as schematized in the left-hand side of Figure 2.3: as working memory scores increase, DD scores should decrease. Statistically speaking, the capacity-based theory predicts that working memory capacity should be a significant predictor of DD scores (e.g., using a standard linear regression), such that the line of best fit derived for the relationship should (i) have a negative slope, and (ii) account for a relatively large portion of the variance in the sample; i.e., measures of goodness of fit such as R^2 should be relatively large. On the other hand, grammatical theories predict no relationship between variation in DD scores and variation in working memory scores, as schematized in the right-hand side of Figure 2.3. Statistically speaking,

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

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.

For space reasons, we will only discuss the results of the comparison between Magnitude Estimation-based DD scores and the serial recall task (experiment 2 from Sprouse *et al*. 2012). The interested reader should see Sprouse *et al*. (2012) for analyses of the Likert scale and *n*-back results, as well as non-DD-based approaches to the statistical investigation of the relationship between the superadditive component of island effects and performance on working memory tasks, all of which yield the same conclusion.

6.1 Participants

The participants in this experiment were 176 self-reported monolingual native speakers of English (152 female), all University of California Irvine undergraduates, who received 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. The analysis below was run on the remaining 173 participants.

6.2 The acceptability rating task

Four island types (*whether* islands, Complex NP islands, Subject islands, and Adjunct islands) were tested, each using a 2 × 2 manipulation of extraction and structural environment as discussed in section 3, yielding a total of sixteen critical conditions. Eight additional sentence types were included to add some variety to the materials, for a total of twenty-four sentence types. Sixteen lexicalizations of each sentence type were created, and distributed among four lists using a Latin Square procedure. This meant that each list consisted of four tokens per sentence type, for a total of ninety-six items. Two orders for each of the four lists were created by pseudorandomizing the items such that related sentence types were never presented successively. This resulted in eight different surveys. The standard was identical for all eight 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 *whether* islands are in (4) above; examples for the other island types are as follows:

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

(5) 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?

(6) 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?

(7) 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. The experiment began with a practice phase during which participants estimated the lengths of seven 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, ten 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 nine items of the survey were practice items (three 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. Participants were under no time constraints during their visit.

6.3 The serial recall task

The serial recall task used eight disyllabic words that were matched for orthographic and phonetic form (CVCVC), approximate frequency, neighborhood density, and phonotactic probability. The eight words were: *bagel*, *humor*, *level*, *magic*, *novel*, *topic*, *tulip*, *woman*. They were recorded by a female native speaker for auditory presentation to the participants. We created ten auditory lists, each containing six of the eight words in a different order. The small pool of eight words was used in each list to prevent the use of mnemonics during the memorization stage (Cowan 2001), whereas the variation created by choosing six for each list added some novelty for the participants. Each participant was presented with all ten sequences in the same order. The words in each list were presented sequentially with an interstimulus interval (ISI) of 500 ms.

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

Table 2.2 *Means and standard deviations of z-scored magnitude estimation scores for each condition (n = 173)*

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

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 2001). 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. Participants were instructed to leave a position blank if they could not recall the correct word so that the standard scoring procedure for serial recall tasks could be used: First, within each trial, a response was counted as correct only if it appeared in the correct position in the response list (1–6). Second, within each position across trials the total number of correct responses was 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 was summed to derive a memory span score (between 0 and 6) for each participant.

6.4 Results

Acceptability judgments from each participant were z-score transformed prior to analysis. The z-score transformation eliminates the influence of scale bias on the size of the DD scores, and therefore increases the likelihood of finding a significant relationship between working memory capacity and DD scores. (See Table 2.2.)

6.4.1 The basic island effects The first question one can ask is whether the basic island effects arise in this sample. Linear mixed effects models revealed a significant main effect of DEPENDENCY, a significant main effect of STRUCTURE, and a significant (superadditive) interaction for each island type (see Table 2.3). Because the interactions are superadditive, pairwise comparisons were used to isolate each of the potential processing costs rather than the

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

Table 2.3 *Two-way linear mixed effects models for each island type and pairwise comparisons for the effects of each structural manipulation (n = 173)*

	whether	complex NP	subject	adjunct
Main effect of DEPENDENCY	.0001	.0001	.0001	.0001
Main effect of STRUCTURE	.0001	.0001	.0001	.0001
DEPENDENCY X STRUCTURE	.0001	.0001	.0001	.0001
Pairwise comparison: DEPENDENCY	.0001	.0001	.0001	.0010
Pairwise comparison: STRUCTURE	.0001	.2260	.3514	.0001

main effects. This is because the interaction (i.e., the extreme unacceptability of the embedded, island condition) could be driving one or both of the main effects. In the pairwise comparisons, the length cost was isolated with a pairwise comparison of the matrix, non-island (4a) and embedded, non-island (4b) conditions. The structure cost was isolated with a pairwise comparison of the matrix, non-island (4a) and matrix, island (4c) conditions. As Table 2.3 indicates, the isolated effect of DEPENDENCY was significant for every island type, as expected. However, the isolated effect of STRUCTURE was not significant for complex NP and subject islands (even with the extremely large sample size of 173). This raises an interesting question of how island effects (the interaction) could be caused by the combination of two processing costs when the cost associated with island structures is only reliably present in the *whether* island, and is reliably absent in the complex NP island and the corrected subject island design (see Sprouse *et al*. (2012) for a more detailed discussion of the independent motivation of each of the components of the resource capacity theory).

6.4.2 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.

Simple linear regressions were performed for each island type using DD scores as the dependent variable, and serial recall scores as the independent variable (see Figure 2.4 and Table 2.4). Two sets of simple linear regressions were run for each island type using the serial recall and DD scores. The first set of regressions was run on the complete set of DD scores for each island type. The second set of linear regressions was run on only the DD scores that were greater than zero for each island type. The logic behind the second analysis is that DD scores below 0 are indicative of a *subadditive* interaction. Neither

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

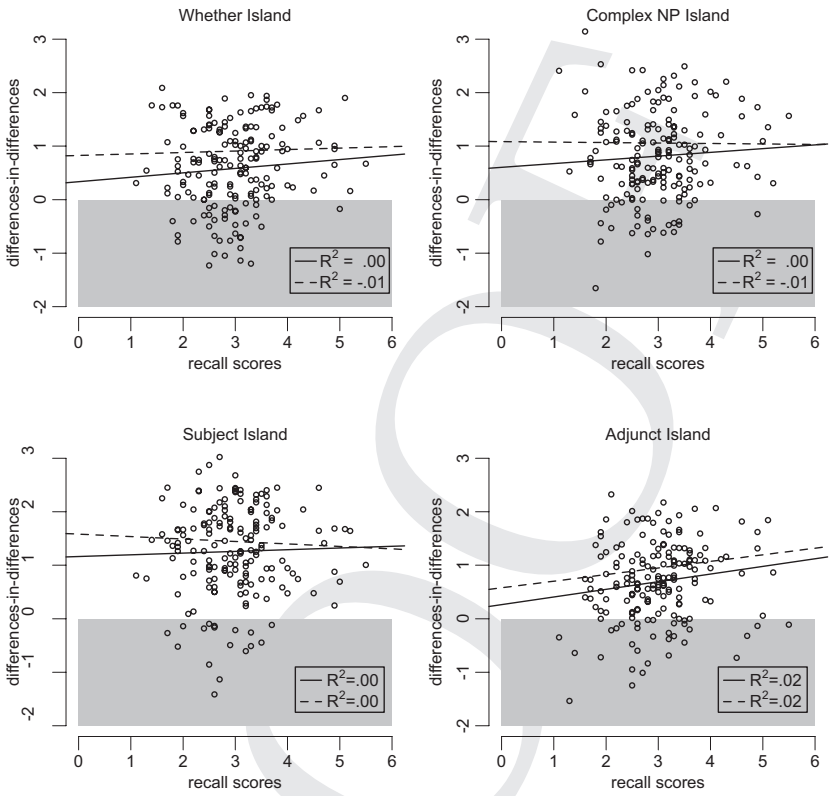

Figure 2.4 Differences-in-differences scores plotted as a function of serial recall scores (n = 173). The solid line represents the line of best fit for all of the DD scores. The dashed line represents the line of best fit when DD scores below 0 are removed from the analysis (shaded grey). Trend lines were fitted using a least-squares procedure. Adjusted R2 for each trend line is reported in the legend.

theory predicts the existence of subadditive effects, which suggests that DD scores below 0 may reflect a type of noise that we may not want to influence the linear regression. By eliminating these potentially unrepresentative scores from the analysis, we increase the likelihood of finding a significant trend in the data.

A simple linear regression analysis finds the line that minimizes the vertical distance between all of the points and the line itself, and reports the coefficients of that line: its intercept with the y-axis and the slope associated with a

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

Table 2.4 *Linear regression modeling differences-in-differences scores as a function of serial recall scores (n = 173)*

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

one-unit change in the predictor variable, which in this case corresponds to recall scores. As with all modeling procedures, a line is always returned by the least-squares procedure, so the first question is whether this line explains the data significantly better than other possible lines, such as a line chosen at random, or a horizontal line that uses the mean as a y-intercept. The adjusted R^2 statistic is a direct measure of the goodness of fit of the line: adjusted R^2 describes the proportion of the variance in the data captured by the line (between 0 and 1) with a slight adjustment based on the number of degrees of freedom in the model. As Table 2.4 reports, six of the eight models had adjusted R^2 of 0 or below, suggesting that they do not account for any of the variance in their respective data sets.[3] The two remaining models only captured 2 percent of the variance in their data sets. As a point of comparison, the line of best fit in the graph in Figure 2.3 that we used to illustrate the prediction of the resource capacity theory has an R^2 of .5 (i.e., 50% of the variance in the data is explained by the line), which within the psycholinguistics literature is generally considered to be a highly meaningful correlation. Unlike *p*-values, there are no broadly agreed-upon conventions for interpreting R^2 values; however, it is safe to assume that the extremely small R^2 values found for each of the island types (even after removing noisy DD scores) are not at all what one would predict for a theory like the resource capacity theory, which relies heavily on a single factor for its explanatory power. These results strongly suggest that there is no evidence of a

[3] Note that negative values are possible with adjusted R^2 because the adjustment for the number of degrees of freedom has the effect of lowering the standard R^2 value slightly. This adjustment is recommended because the standard R^2 value is a biased (inflated) statistic. Negative values suggest that the biased (inflated) R^2 statistic was at or near zero, and that the degrees-of-freedom correction brought the statistic below zero.

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

relationship between DD scores and recall scores, contrary to the predictions of the resource capacity theory, and consistent with the predictions of grammatical theories.

7 Moving forward with the debate

This chapter began with three primary goals: (1) to illustrate the fundamental differences between grammatical explanations and reductionist (i.e., "processing") explanations of island effects, (2) to establish the empirical facts that must be explained by any theory of island effects, and (3) to discuss methods of evaluating the empirical adequacy of reductionist explanations. The first step was to establish the relationship between grammatical theories and parsing theories as two related descriptions of the human sentence-processing system. In the process, it became clear that in order for reductionist theories to be distinct from grammatical theories, reductionist theories must rely on the non-structure-building components of the sentence-processing faculty. We proposed the following formalization of terms: (i) grammatical theories posit that the required structure-building operations are illegitimate, while (ii) reductionist theories posit that the structure-building operations are possible, but aren't carried out in specific circumstances due to constraints on the resources available to the parsing system. We also argued that in order to have true explanatory power, reductionist theories must be explicit about the mechanisms that give rise to island effects, as these mechanisms must be independently motivated by constructions other than island effects. Furthermore, these mechanisms must account for the facts of island effects: when tested using a factorial design, the low acceptability reported for island violation sentences is not a simple linear sum of the factors, but rather a superadditive interaction. This superadditive effect requires additional assumptions beyond the individual mechanisms of a linear reductionist theory, such as the limited resource capacity assumptions proposed by Kluender and Kutas (1993b).

Once the mechanisms and assumptions of the reductionist theory are laid out, there are basically two possible approaches to testing the theories: (i) test that the mechanisms and assumptions are independently motivated (i.e., required by constructions other than island effects), and (ii) test any novel predictions of the extra mechanisms and assumptions postulated by the reductionist theory. In this chapter, we focused on strategy (ii) by reporting one of the experiments from Sprouse *et al*. (2012) that tested a possible prediction of the resource capacity theory of Kluender and Kutas (1993b): that the superadditive component of island effects would be inversely proportional to the working memory capacity of participants. The results of that study suggest that the prediction of the resource capacity theory does not hold; however, it is at least logically possible to construct other types of reductionist theories. Though doing so is well beyond

PROOF

Trim: 228mm × 152mm Top: 12.548 mm Gutter: 18.98 mm
CUUK2307-02 CUUK2307/Sprouse ISBN: 978 1 107 00870 0 May 3, 2013 20:2

the scope of this chapter, it is our hope that this chapter lays out a framework for evaluating the logical and empirical adequacy of reductionist theories, and provides a clear discussion of exactly what is at stake in the debate between grammatical and reductionist approaches to island effects.

PROOF