Experimental syntax and the cross-linguistic variation of island effects in English and Italian

Jon Sprouse (corresponding author)
Department of Cognitive Sciences
University of California, Irvine
jsprouse@uci.edu

Ivano Caponigro
Department of Linguistics
University of California, San Diego
ivano@ucsd.edu

Ciro Greco
Department of Psychology
University of Milan-Bicocca
c.greco12@campus.unimib.it

Carlo Cecchetto
Department of Psychology
University of Milan-Bicocca
carlo.cecchetto@unimib.it

*This work was supported in part by National Science Foundation grant BCS-0843896 to JS. We would like to thank Michela Marchesi for assistance collecting data for the Italian WH-dependencies experiment. We would also like to thank audiences at Harvard University and Johns Hopkins University for helpful comments on earlier versions of this work. All errors remain our own.
Experimental syntax and the cross-linguistic variation of island effects in English and Italian

Abstract

Cross-linguistic variation is critical to (at least) two debates at the forefront of linguistic theory: (i) To what extent can complex syntactic phenomena be explained with domain-general cognitive mechanisms rather than specific syntactic mechanisms? and, (ii) To what extent can the acquisition of complex syntactic phenomena be achieved with domain-general learning strategies rather than specific syntactic strategies? Our goal in this article is to apply formal acceptability judgment experiments (experimental syntax) to a high-profile example of cross-linguistic variation (syntactic island effects in English and Italian) in order to demonstrate both the benefits and limits of these experimental techniques in providing additional information that bears on these important questions. In a series of formal experiments, we compare two definitions of island effects (the traditional definition and the factorial definition made available by formal experiments), two types of long-distance dependencies (WH-dependencies and RC-dependencies), two languages (English and Italian), and four types of syntactic islands (\textit{wh/whether}, complex NP, subject, and adjunct). The results reveal a finer grained pattern of island effects across languages and dependency types than was previously revealed with traditional methods. These results have consequences both for debates about the role of domain-general mechanisms in the existence of island effects, and for debates about the mechanisms underlying language acquisition. These findings also suggest that formal acceptability judgment experiments, and crucially the factorial definitions that they enable, can play a substantive role in future cross-linguistic studies.

1. Introduction

It is well-known that acceptability judgments form the primary empirical foundation of syntactic theories (Chomsky 1965, Schütze 1996), and that the vast majority of the acceptability judgments published in the syntactic literature are collected using relatively informal methods. In recent years there has been a growing interest in the use of more formal methods for the collection of acceptability judgments, both due to concerns regarding the reliability of informally collected judgments (e.g., Ferreira 2005, Wasow and Arnold 2005, Gibson and Fedorenko 2010), and due to indications that more formal methods may reveal previously unobserved patterns in the data (e.g., Keller 2000, Featherston 2005, Sprouse et al. 2011). While concerns about the reliability of informally collected judgments has turned out to be grossly overstated (Sprouse and Almeida \textit{in press} a, Sprouse, Schütze, and Almeida \textit{submitted}, Sprouse and Almeida \textit{in press} b), it is still an open question to what extent formal methods may reveal previously unobserved patterns, especially with respect to the cross-linguistic variation of complex syntactic phenomena. As such, our goal in this article is to apply formal experimental methods, commonly known as experimental syntax techniques, to a high-profile example of cross-linguistic variation: syntactic island effects in English and Italian (e.g., Rizzi 1982). It is our hope that the systematic application of experimental syntax techniques, and crucially the quantitative definitions that they make available, to this longstanding research topic will demonstrate both the benefits and limits of applying formal methods to cross-linguistic syntactic research. Crucially, our goal is not to
provide a single theoretical analysis of the results that we observe within a specific grammatical framework, as that would be a relatively small (albeit important) contribution. Instead, our goal is to demonstrate the potential ramifications of the results for two of the larger debates at the forefront of both syntactic theory and cognitive science more generally: (i) To what extent can complex syntactic phenomena be explained with domain-general cognitive mechanisms rather than specific syntactic mechanisms?, and (ii) To what extent can the acquisition of complex syntactic phenomena be achieved with domain-general learning strategies rather than specific syntactic strategies. As we will see, cross-linguistic data is critical to progress on these larger issues, therefore it is in the best of interest of the field to apply any potential methodological advances to these questions if there is any possibility of refining the empirical foundation of the theory.

This article is organized as follows. In Section 2, we present our reasons for choosing island effects as a case study in the use of formal acceptability judgment experiments for studying cross-linguistic syntactic variation. We also provide a brief review of current beliefs about the variation in island effects between English and Italian. In Section 3, we discuss two potential experimental definitions of island effects: the traditional definition that has been used in (most, if not all) previous studies, and the factorial definition that has been recently developed to better control for domain-general (reductionist) approaches to island effects (Sprouse 2007, Sprouse et al. 2011, Sprouse et al. 2012). We also discuss the overall design of the present study: a series of formal acceptability judgment experiments that investigate the variation between both languages (English and Italian), across two construction types (wh-interrogatives and relative clauses), and across four island types (whether/wh, complex NP, subject, and adjunct). Sections 4 through 7 present the specific designs, results, and statistical analyses of the formal acceptability judgment experiments. Section 8 provides a discussion of the results and their consequences for: (i) syntactic theories, (ii) the debate about the complexity of grammars, and (iii) the debate about the complexity of learning strategies. Section 9 concludes with a brief discussion of the implications of this study for the future use of formal methods in cross-linguistic syntactic research.

2. Island effects and cross-linguistic variation

We have chosen island effects as our case study in this investigation because we believe the role that island effects play in the two larger debates about the complexity of grammars and the complexity of learning strategies is relatively clear (e.g., Sprouse et al. 2012 and Pearl and Sprouse to appear). In this section we will review these issues in some detail to provide a common background for discussing the ramifications of our experimental results. However, we hope it is clear throughout that island effects are simply a first case study – similar consequences would hold for any complex syntactic phenomenon that has been observed to vary in a potentially constrained manner cross-linguistically.

2.1 An introduction to island effects

One of the defining characteristics of human language is the existence of long-distance dependencies between two (or more) elements in a sentence. For example, the wh-interrogative clauses in (1) illustrate a long-distance dependency between the wh-word or wh-phrase at the beginning of the sentence, which is often called the antecedent or the filler, and the argument
position of an embedded verb, which is often called the *gap position*. Although long-distance dependencies are unconstrained with respect to length as measured in number of words or number of clauses, as in (1), there do appear to be constraints on the types of structures that can contain the gap position, as in (2). In the examples below and throughout the paper, the antecedent will be in italics and the gap position will be indicated with underscores.

(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]?*

Following Ross (1967), the unacceptability that arises when the gap position occurs inside one of the prohibited structures in (2) is often referred to as an *island effect*, which draws on the metaphor that the prohibited structures are *islands* that prevent the wh-words or wh-phrases from *moving* to the front of the sentence. Though island effects are typically exemplified by *WH-depencies*, i.e. dependencies between a wh-word or phrase and a gap inside an interrogative clause as in (2), the same effects crucially arise with several different types of long-distance dependencies in human languages: for instance dependencies between the head introducing a headed relative clause and the gap within the relative clause, henceforth **RC-dependencies** (3), dependencies between the topIALIZED pre-posed constituent and the lower gap (4), and dependencies between the preposed adjective and the lower gap in adjective-though constructions (5). All the b. examples in (3)-(5) exemplify an extraction out of a *whether*-island.

(3) **RELATIVE CLAUSE FORMATION (RC-DEPENDENCIES)**
   a. I like the *car* that you think [that John bought ___].¹
   b. *I like the *car* that you wonder [whether John bought ___].

¹ We remain neutral regarding the analysis of headed relative clauses: the filler of the gap could either be the head noun (as in raising analyses: e.g., Vergnaud 1974, Kayne 1994, Bianchi 1999, Bhatt 2002, Donati and Cecchetto 2011), or a (potentially null) relative pronoun (e.g., Chomsky 1981, Browning 1986).
(4) **TOPICALIZATION**
   a. I don’t know who bought most of these cars, but *that car*, I think [that John bought __].
   b. *I know who bought most of these cars, but *that car*, I think [whether John bought__].

(5) **ADJECTIVE-THOUGH CONSTRUCTIONS**
   a. Smart though I think [that John is __], I don’t trust him to do simple math.
   b. *Smart though I wonder [whether John is __], I trust him to do simple math.

As we will see in Section 2.4, the variation in island effects between English and Italian involves both interrogative clause formation and related WH-dependencies (2) and headed relative clause formation and related RC-dependencies (3).

2.2 Cross-linguistic variation and the complexity of grammars

The existence of island effects has led many researchers to posit complex grammatical constraints, often known as *island constraints*, on the formation of long-distance dependencies. Although syntactic constraints are by far the most common grammatical approach to island effects (e.g., Ross 1967, Chomsky 1973, Huang, 1982, Chomsky 1986), there are also semantic approaches (e.g., Szabolcsi and Zwarts 1993, Truswell 2007, and Abrusán 2011), and pragmatic approaches (e.g., Kuno 1976, Erteschik-Shir 1979, Kuno 1987, Kuno & Takami 1993, Goldberg 2007). Grammatical approaches, whether they be syntactic, semantic, or pragmatic in nature, add a layer of complexity to the grammar that has far-reaching consequences for both the architecture of the system (i.e., there must be both one or more mechanisms that give rise to long-distance dependencies and one or more constraints on the application of those mechanisms), and for the complexity of the learning problem faced by children acquiring the language (which we will discuss in more detail in Section 2.3).

Given the far reaching consequences of the grammatical approach to island effects, it is perhaps unsurprising that there is a second class of theories – known as *reductionist* theories – that explicitly reject the grammatical approach. 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, Gieselmann et al. *to appear*).

According to reductionist approaches, the perception of unacceptability that characterizes island effects is not due to the ungrammaticality of the sentence, because in fact, such sentences are grammatical. Instead, the unacceptability arises as a by-product of the processing requirements of the sentence. Reductionism is a strong theme in every science (often under the rubric of *Occam’s razor*), including all grammatical theories that we are aware of (it is even encoded in the name of one generative theory – *minimalism*). As such the goal of reductionist approaches to simplify the quantity and quality of grammatical mechanisms is an important one to consider to the extent that the empirical facts license it.

Cross-linguistic variation plays a central role in evaluating the relative empirical adequacy of grammatical and reductionist approaches to any syntactic phenomenon. This is because one of the strengths of grammars is their ability to capture cross-linguistic variation; indeed, in many linguistic theories grammars are the seat of all cross-linguistic variation. If a given phenomenon does indeed exhibit cross-linguistic variation, then reductionist approaches to that phenomenon must capture that variation in order to satisfy the *all else being equal* clause of
rubric’s like Occam’s razor (i.e., a simpler theory is to be preferred all else being equal). In other words, reductionism is only to be favored if the reductionist theory captures the same set of empirical facts as its competitor theory. Cross-linguistic variation is potentially problematic for reductionist approaches that are in competition with grammatical approaches, because by definition, such reductionist approaches may only invoke extra-grammatical cognitive mechanisms to account for the phenomenon, and extra-grammatical cognitive mechanisms tend to be relatively universal across human cultures.

As a concrete example, a popular reductionist approach to island effects posits that island effects arise as the result of limited the working memory capacity that is available to parse a sentence that involves both a long-distance dependency and the syntactic structures that characterize island effects. Cross-linguistic variation in the presence of island effects is difficult (although not necessarily impossible) to capture under this approach, because domain-general cognitive mechanisms like working memory capacity likely do not vary among humans according to the boundaries of different languages (i.e., English speakers likely do not have less working memory than Italian speakers). The problem is compounded if the cross-linguistic variation is constrained – i.e., if only a subset of the logically possible variations arise in human languages – as this suggests a complex constraint on the range of variation of a domain-general cognitive mechanism like working memory capacity. Although it is not impossible in principle to capture (constrained) variation under a reductionist approach (e.g., tying the working memory cost to relatively superficial morphosyntactic properties of the sentence), grammatical approaches may be better suited for the task. Given this tension between grammatical approaches, reductionist approaches, and cross-linguistic variation, it is of paramount importance for the field to investigate any potential empirical advance (such as the use of formal acceptability judgment experiments) in the domain of cross-linguistic syntax.

2.3 Cross-linguistic variation and the complexity of learning strategies

As briefly mentioned in the previous section, grammatical approaches to island effects not only increase the complexity of the grammatical architecture, but also potentially increase the complexity of the learning problem faced by children acquiring the language: learning a complex grammatical constraint is necessarily more complex than not learning such a constraint. This increased complexity can be neutralized by postulating innate, domain-specific learning biases as part of the genetic endowment of human children - an approach that is often called the Universal Grammar (UG) approach (Chomsky 1965). Although the UG approach successfully counteracts the added complexity of grammatical approaches to island effects, it comes with its own cost. Much like reductionist approaches to syntactic phenomena that emphasize domain-general cognitive mechanisms are generally preferred over grammatical approaches, reductionist approaches to language acquisition that emphasize domain-general learning mechanisms are generally preferred over UG approaches.

Recent computational modeling work has suggested that island effects in English can in principle be learned from realistic child-directed input with relatively few (if any) UG-based mechanisms (Pearl and Sprouse to appear). Although this new model may appear to be an empirical argument in favor of a reductionist approach to acquisition, by relying extensively on domain-general learning mechanisms, it makes very specific predictions about the nature of cross-linguistic variability. By eschewing (most, if not all) UG-based mechanisms, the model primarily learns the pattern of island effects for a given language from the input that it is
presented. There are no additional constraints on the possible patterns of island effects imposed by the learning mechanism. What this means in practice is that this model predicts no constraints on the variation of island effects cross-linguistically: any potential pattern of results can be derived given the corresponding input.

The problem posed by constrained variation in island effects for this model is straightforward. If there is indeed constrained variation in island effects cross-linguistically, then this model would force us to conclude that the apparent constraint is simply a coincidence: the inputs of the languages in question just happened to not include the information that would be necessary to lead to the unobserved patterns of island effects. In contrast, a learning theory that includes UG-based mechanisms could impose constraints on the learners such that the number of potential outcomes (i.e., patterns of island effects) would be limited. Given this tension between UG-based approaches to acquisition, reductionist approaches to acquisition, and cross-linguistic variation, it is of paramount importance for the field to investigate any potential empirical advance (such as the use of formal acceptability judgment experiments) in the domain of cross-linguistic syntax.

2.4 Previous observations of the cross-linguistic variation of island effects in English and Italian

At this point it may be useful to move beyond abstract references to cross-linguistic variation, and instead review previous work on the cross-linguistic variation in island effects between English and Italian. This will provide a foundation for both the evaluation of the design of our experiments, as well as the consequences of the results. The seminal study on island effects in Italian is by Rizzi (1982), who first observed that Italian does not exhibit the same set of island effects as English. Specifically, Rizzi observed that whereas English exhibits wh, complex NP, and subject islands as in (2) above, Italian only appears to exhibit complex NP islands (all of the judgments about Italian in the examples below are from Rizzi 1982):

(6) **WH ISLAND** (Rizzi 1982: 50, ex. 6)

\[
\text{Tuo fratello, a cui mi domando che storie abbiano raccontato, was very worried}
\]

\[
\text{'Your brother, who I wonder what stories they told, was really worried.'}
\]

(7) **COMPLEX NP ISLAND** (Rizzi 1982: 51, ex. 9a)

\[
\text{*Questo incarico, che non sapevo la novit\`a che avrebbero affidato a te,\ldots this task that not knew the news that they may have assigned to you}\\
\text{(*'This task, which I didn’t know the news that they may have assigned to you \ldots’)}
\]

---

2 Here and throughout the paper, italics are ours. Also, both the head and the relative pronoun are italicized when they both occur as a way of being non-committal about which of them is the actual antecedent of the gap. Glosses and translations have been modified or added, if missing in the original examples. Translations are in parentheses when they translate unacceptable sentences and are preceded by a “*” when they are unacceptable strings in English.
Rizzi (1982) did not directly investigate adjunct islands. However, for completeness we can anticipate the results of the current experiments slightly and note that Italian also appears to exhibit adjunct islands (see also Stepanov 2007 for a broad review of the languages that demonstrate adjunct islands):

One fact worth noting at this point is that Rizzi (1982) focused on relative clause formation as opposed to wh-interrogative clause formation in his study of island effects in Italian. This contrasts with most studies of island effects in English, which have tended to focus on wh-interrogatives rather than relative clauses. Rizzi (1982) offers a principled explanation for this choice, at least with respect to wh-islands: whereas Italian wh-interrogatives appear to demonstrate wh-islands as illustrated in (10), there is reason to believe that sentences with two or more wh-words in a single clause would be ruled out independently in Italian. Sentences like (11) are unacceptable in Italian, despite no obvious island violation, presumably because there are two wh-words (originating) in a single clause (in other terms, Italian does not allow wh-in-situ even when a second wh-phrase moves to the dedicated position in the left periphery, nor does it allow for more than one wh-phrase in the left periphery of the same clause):³

³We are reporting here Rizzi’s judgments from the late 70s. Today, interrogative clauses like Chi ha comprato che cosa? (‘Who bought what?’) are acceptable in many varieties of Italian, for examples in the journalistic jargon, possibly as a borrowing from English. If a syntactic transfer is taking place, it is still on-going. For example, Chi ha comprato che cosa? sounds better than other combinations of wh-phrases (including those that are acceptable in English). At the present time, the distribution of different wh-in-situ phrases in Italian displays a complex pattern (see Moro 2011).
Although it is not directly discussed by Rizzi (1982), the fact that wh-island effects with wh-interrogatives are only *apparent* is a critical component of many syntactic theories of island effects. This is because many theories relate wh-interrogative and relative clause formation through a single grammatical operation (e.g., movement), and capture the existence of island effects through a constraint (or constraints) on that operation. In fact, it is the existence of island effects for these two types of dependencies that motivates their treatment as a single phenomenon (e.g., as A'-dependencies in generative theories). If Italian were indeed to exhibit two different sets of island effects for the two different constructions, then this would suggest that the theory must be more complicated. Rizzi’s (1982) explanation of the apparent existence of wh-islands with wh-interrogatives in terms of a prohibition on multiple wh-words (or phrases) in Italian maintains the simpler version of the theory.

There is a potential problem with Rizzi’s (1982) explanation for island effects in Italian wh-interrogatives: not every island type involves a wh-word (or phrase). For any island type that does not involve a wh-word (or phrase) as part of its structural definition, it is possible to construct wh-interrogatives that involve that island type without running afoul of the prohibition on multiple wh-words (or phrases). If these sentences are indeed unacceptable, then that unacceptability either suggests a true island effect, or requires an additional explanation beyond the prohibition on multiple wh-words (or phrases). As demonstrated in (2a), it is possible to construct a structure that is very similar to a wh-island that involves an embedded interrogative without using a true wh-word by substituting the complementizer *whether* (or *if*) for the wh-word in the specifier position of the embedded CP. The other examples in (2) demonstrate that most island types do not involve wh-words or phrases therefore can be investigated in Italian in wh-interrogatives without modification. Because Italian raises the possibility of variation between wh-interrogative and relative clause formation (cross-construction variation), we will investigate both constructions in this study. As already introduced above, we will refer to wh-interrogative formation as WH-dependencies and relative clause formation as RC-dependencies, just for ease of exposition, without any commitment about the similarities and differences between the two constructions (including the possibility that overt or covert wh-words or phrases may be used to form relative clauses as well).

As per the discussion in sections 2.2 and 2.3, Rizzi’s (1982) cross-linguistic observations have profound consequences for both theories of grammar and theories of acquisition. The cross-linguistic, and potentially cross-construction, variation that Rizzi observed suggest that island effects may be best captured by a grammatical constraint rather than as a consequence of domain-general cognitive mechanisms (which are less likely to vary according to linguistic populations or constructions). This receives support from Torrego’s (1984) observation that Spanish WH-dependencies exhibit the same pattern of island effects as Italian: complex NP and adjunct island effects are present, but wh-islands and subject island effects are absent. Torrego’s results also raise the possibility that the cross-linguistic variation in island effects is constrained such that any language that exhibits wh-islands will also exhibit subject islands (and vice versa), and any language that that fails to exhibit wh-islands will also fail to exhibit subject islands (and vice versa). Although the scope of our investigation extends beyond specific syntactic analyses, Rizzi’s (1982) proposal captures this constrained variation directly. Rizzi proposes adopting Chomsky’s (1973) Subjacency Condition to capture island effects. The Subjacency Condition states that the grammatical operation *movement*, which relates the antecedent and the gap in WH-dependencies and RC-dependencies cannot cross two or more bounding nodes. Rizzi proposes defining the set of possible bounding nodes such that languages...
like English, which display both wh-islands and subject islands, have one set of bounding nodes (in more modern terms, TP and DP), and that languages like Italian and Spanish, which do not display either wh-islands or subject islands, have a different set of bounding nodes (CP and DP). This analysis effectively captures both the cross-linguistic variation and the apparent relationship between wh-islands and subject islands.

Rizzi’s (1982) Subjacency analysis also lends itself to the second consequence of constrained variation: a constraint on the learning mechanism that prevents the unobserved pattern from arising regardless of the input available to children. Under Rizzi’s approach, the Subjacency Condition is an innately specified principle that is available to the acquisition system. The set of possible bounding nodes is also innately specified, such that the child needs only choose which of the possible bounding nodes is used in their particular language. Of course, there are many possible analyses of these facts, and many possible learning mechanisms that could give rise to this particular pattern of island effects; we mention the Subjacency analysis here simply to illustrate the consequences that Rizzi’s cross-linguistic observations have for theories of grammar and acquisition. Given these consequences, we decided to investigate both WH-dependencies and RC-dependencies in both English and Italian for four different island types: wh/whether (depending on the dependency type to avoid the multiple-wh prohibition in Italian), complex NP, subject, and adjunct islands.

3. The definition of island effects

One of the primary benefits of formal acceptability judgment experiments is that they allow us to leverage complex quantitative definitions of syntactic phenomena in order to control for multiple sources of unacceptability (Sprouse 2007, Sprouse et al. 2012). In this section we compare and contrast the traditional definition of island effects that has been used by Rizzi (1982) and others with the factorial definition of island effects that has recently been developed for use in formal acceptability judgment experiments (Sprouse 2007, Sprouse et al. 2011, Sprouse et al. 2012), paying particular attention to the ability of the factorial definition to provide information relevant to the debate between grammatical and reductionist approaches to island effects. As our experiments rely on this factorial definition, we also take this opportunity to present a comprehensive view of the logic and design of the present study.

3.1 The traditional definition of island effects

The traditional definition of island effects employed by Rizzi (1982) and many others is relatively intuitive: if a given sentence is judged sufficiently unacceptable, then that unacceptability is need of an explanation. If multiple sentences can be constructed using different lexical items but the same structural configuration, then the lexical items can be eliminated as a potential source of the unacceptability. Although it is difficult in practice to dissociate syntactic and semantic properties because of their close relationship, to the extent that the intended meaning of the sentences is both plausible and well-formed according to existing semantic theories, then semantics can be (at least partially) eliminated as a potential source of the unacceptability. Similar elimination procedures can be tested for potential prosodic or pragmatic explanations. If all of these elimination procedures are successful, then this generally leaves either a syntactic explanation or a reductionist explanation based upon non-syntax-related mechanisms of the sentence processing system. Although the inability to tease apart syntactic
and reductionist explanations under the simple traditional definition was one motivation for the development of the factorial definition (which we discuss in the next section), it is not the case that evidence relevant to the debate was impossible to collect prior to the development of the factorial definition. As previously mentioned, cross-linguistic variation is one major piece of evidence relevant to the debate (see also Phillips to appear for more discussion). Some researchers have even attempted to demonstrate that it is possible to manipulate the structure of a sentence in a way that manipulates the presence of a traditionally defined island effect without any obvious manipulation of the processing requirements of the sentence (e.g., Dillon and Hornstein to appear). There are also numerous sentence processing studies that have attempted to measure the effect of island effects in real-time sentence processing (e.g., Phillips 2006 and reference therein, Hofmeister and Sag 2010, Yoshida et al. to appear). Any comprehensive theory of island effects must ultimately explain all of these data.

One of our main goals in the present study is to compare both the traditional and factorial definition of island effects. Therefore, before introducing the factorial definition, it is important to take a moment to discuss precisely how the traditional definition can be translated into a formal acceptability judgment experiment. The traditional definition involves rating the critical island effect sentences and judging whether they are sufficiently unacceptable to merit an explanation. To translate this into a formal experiment, we simply need to quantitatively define sufficiently unacceptable. Although different researchers may have different criteria for this, we believe that the experimental literature can offer a principled definition. Cohen (1988, 1992) observes that differences between means of conditions that are 0.5 standard deviations or larger are generally visible to the naked eye of trained researchers without the need to run statistical tests (he called these medium effect sizes). It seems reasonable to conclude from the discussion of island effects in the syntactic literature that linguists view island effects as phenomena that are visible to the naked eye without the need to run statistical tests. This suggests that island effects must be at least 0.5 standard deviations less acceptable than the least acceptable sentence that clearly does not require an explanation (so that statistical tests would not be necessary to identify them). To be as conservative as possible, we can codify this as the requirement that island effects be at least 0.5 standard deviations below the middle of the range of acceptability.

This definition is relatively straightforward to assess in formal judgment experiments because it is common practice to z-score transform judgment ratings prior to analysis. The z-score transformation eliminates several types of scale bias that can arise between participants (e.g., using one of the scale, or using a limited range of responses) by converting each participant’s ratings to a standard scale: each participant’s mean rating is subtracted from each of the participant’s ratings, and the resulting difference is divided by the standard deviation of the participant’s ratings. This process has two effects: the participant’s mean rating is transformed to 0, and each rating is transformed onto a scale that represents the number of standard deviations that rating is away from the mean (i.e., a rating of 0.5 is 0.5 standard deviations from the mean). By ensuring that the items in the survey are balanced both in location and number across the range of acceptability, the mean rating of each participant should correspond closely to the middle value of acceptability for that participant. By averaging all of the participant’s z-score transformed ratings together, the 0 value of the average should correspond closely to the middle value of acceptability for the sample. This means that the criterion for establishing island effects under the traditional definition is a rating of -0.5 on the z-score transformed scale of acceptability. We will use this criterion as a quantitative equivalent of the traditional definition in this study.
because we believe it is both as principled and as conservative as we can be; however, readers are encouraged to use their own definition in interpreting the results.

3.2 The factorial definition of island effects

Sentences that give rise to island effects generally contain two specific components: (i) a long-distance (often bi-clausal) dependency (e.g., a WH-dependency or a RC-dependency), and (ii) a complex syntactic structure that we call an island structure. What is interesting about these two facts is that they can be used to motivate the simplest possible reductionist theory: if there is a processing cost associated with the operations necessary to build long-distance dependencies, and if there is a processing cost associated with the operations necessary to build the island structures, then these two costs could combine to create the unacceptability that we call island effects without the need for any grammatical constraints. The factorial definition of island effects is explicitly designed to control for these two potential processing costs. First, we can isolate the effect of dependency length on acceptability by contrasting a sentence with a short dependency (i.e., both the antecedent and the gap are in the same clause) as in (12a), with a sentence that contains a longer dependency (i.e., the antecedent and the gap are in two different clauses) as in (12b). Similarly, we can isolate the effect of processing island structures by contrasting a sentence with an island structure as in (12c) with a sentence that does not contain an island structure as in (12a). Finally, we can measure the effect on acceptability of processing both long-distance dependencies and island structures -- the island effect itself -- by combining both in a single sentence as in (12d).

(12) A factorial design for measuring island effects: STRUCTURE × 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 (12) indicate, this design contains two factors (STRUCTURE and GAP-POSITION) each with two levels (ISLAND/NON-ISLAND and MATRIX/EMBEDDED) (see also Sprouse 2007, Sprouse et al. 2011, Sprouse et al. 2012).

The simplest reductionist theory predicts that the relationship between the two processing costs should be linearly additive: the cost of processing long-distance dependences [(12a)-(12b)] plus the cost of processing whether clauses [(12a)-(12c)] should equal the cost of performing both together [(12a)-(12d)]. 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 to investigate the whether islands given in (12) in an acceptability judgment experiment (see Section 4 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 effect sentence (condition (12d), which is in the bottom right quadrant of each panel of Figure 1) relative to the baseline condition (condition (12a), which is in the top left quadrant of each panel) is simply the sum of the two main effects. If the simple reductionist theory is true, then the result will resemble the left panel in Figure 1, suggesting that 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.

Results like the one exemplified in the right panel of Figure 1 raise problems for the simplest reductionist theory. The right panel of Figure 1 demonstrates a super-additive interaction: the combined effect of the two costs is greater than the (linear) sum of the individual costs; in other words: \([(12a)-(12b)] + [(12a)-(12c)] < [(12a)-(12d)]\). The term super-additive refers to the fact that the overall difference between the island effect sentence and the basic control is greater than the sum of the component factors. A super-additive 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, super-additivity can be visually identified by non-parallel lines that “open” in the direction of interest. Super-additivity suggests that something beyond the two explicit factors is contributing to the unacceptability of island effects. This extra mechanism could be a grammatical constraint such as the Subjacency Condition, or it could be some mechanism that makes the two factors dependent on each other. For example, Kluender and Kutas (1993) propose that the processes that are deployed to parse WH-dependencies and the processes that are deployed to parse island structures draw from a single pool of working memory resources. This in effect makes the two factors dependent on each other, as an increase in the resources used by one decreases the resources that are available for the other. If that pool of working memory resources is smaller than the amount necessary to successfully deploy both sets of processes, then the super-additive
unacceptability could be the result of over-extending the limited pool of resources (see Sprouse et al. 2012 for a review and test of this proposal). The bottom line is that linearly additive results like those in the left panel of Figure 1 suggest no place for grammatical constraints, as the unacceptability of island effects can be completely explained by the acceptability costs of each of the component factors; super-additive results like those in the right panel of Figure 1 suggest a place for either grammatical constraints or a complex linking assumption between the component factors, because the unacceptability of island effects cannot be completely explained by the acceptability costs of each of the component factors.

Although it should now be clear how the factorial definition provides more information about island effects than the traditional definition (in the form of comparing linear and super-additive effects), a few caveats are in order. First, as discussed in the previous paragraph, the existence of a super-additive interaction is not sufficient to conclude that a grammatical constraint is at work. It is also logically possible for there to be a non-grammatical linking mechanism that gives rise to the interaction, such as a limitation in working memory capacity (e.g., Kluender and Kutas 1993). Testing this latter possibility requires investigating the specific properties of the proposed linking mechanisms, such as the specific components of working memory that are recruited during sentence comprehension (see Sprouse et al. 2012 for a test of the working memory proposal by Kluender and Kutas 1993; see Caplan and Waters 1999 and Cecchetto and Papagno 2011 for two opposing views on the role of working memory during sentence comprehension). Second, because the factorial definition is predicated upon statistical tests (e.g., factorial ANOVAs or linear mixed effects models), it is possible to observe a statistically significant interaction for very small differences between conditions, especially with very large sample sizes. This is because if there is any difference between two conditions, no matter how small (as is universally the case with different sentences), it is always mathematically possible to achieve a $p$-value below .05 if the sample size is large enough (e.g., Shaver 1993, Cohen 1994, Nickerson 2000). This mathematical problem with $p$-values has led many statisticians to draw a distinction between statistical significance and practical significance. Whereas a statistically significant effect is one that reaches a specified $p$-value, a practically significant effect is one that is both statistically significant and meets the researcher’s criteria for theoretical relevance. Theoretical relevance cannot be determined by a statistical test; it can only be determined using a researcher’s scientific knowledge of the theory, phenomenon, and previous results (Shaver 1993, Cohen 1994, Gigerenzer 1994, Nickerson 2000). This means that, much like we had to specify a criterion for the presence of island effects under the traditional definition, we should specify a criterion for what counts as a practically significant interaction under the factorial definition.

One way to define practical significance for the factorial definition is to choose a minimum criterion for the size of the super-additive component of the interaction (i.e., the “extra” bit of unacceptability that isn’t predicted by the component factors). A common and relatively definition of effect size for super-additive interactions is a measure known as a differences-in-differences (DD) score (Maxwell and Delaney 2003). As the name implies, a differences-in-differences score calculates the difference between the differences that are inherent in a factorial design (imagine subtracting the size of the difference between the two leftmost points in one of the figures from the difference between the two rightmost points in the same figure). In our $2\times2$ design, we can calculate a DD score by first calculating the difference ($D_1$) between the NON-ISLAND/EMBEDDED condition and the ISLAND/EMBEDDED condition: $D_1 = (12b) - (12d)$. Next, we can calculate the difference ($D_2$) between the NON-ISLAND/MATRIX condition and the
14

ISLAND/MATRIX condition: $D_2 = (12a) - (12c)$. Finally, we can calculate the difference between these two difference scores: $DD = [(12b) - (12d)] - [(12a) - (12c)]$. We can then use this DD score (which represents the super-additive component of the interaction) to determine whether the interaction is practically significant. As discussed in Section 3.1, island effects are generally considered relatively large effects, such that they should be visible without statistical tests. Therefore we will apply the same logic as we did with traditional definitions and define practically significant interactions as those that have a DD score that is equal to or larger than 0.5 on the z-score transformed scale.

3.3 The logic and design of the present study

Given Rizzi’s (1982) observation that Italian has a different set of island effects than English, we are interested in comparing the pattern of results for the two languages. Furthermore, given Rizzi’s observation that there may be variation in the set of island effects that hold between WH-dependencies and RC-dependencies, we are also interested in comparing the pattern of results for the two construction types in each language. Rizzi’s (1982) initial investigation focused on wh- and subject islands because these are the two that do not appear to arise with Italian RC-dependencies, with complex NP island serving as a control case to demonstrate that at least some island types do appear in Italian. As such, we have chosen to investigate these three island types, i.e. *whether*, complex NP, subject, and one more: adjunct islands. These four islands form a solid foundation for the investigation of cross-linguistic and cross-construction variation in island effects, as they have been previously investigated with respect to both of the larger debates of interest: the grammar versus reductionist debate in syntax (e.g., Sprouse et al. 2012), and the UG versus reductionist debate in acquisition (e.g., Pearl and Sprouse to appear). Finally, we are interested in comparing the traditional definition used in previous studies to the factorial definition that is licensed by formal acceptability judgment experiments. Our goal is to both replicate the previously reported traditional results under formal experimental conditions, as well as investigate to what extent the factorial definition may reveal additional detail about the pattern of results (i.e., linear additivity versus super-additivity).

In order to accommodate all of these experimental factors, we conducted 8 formal acceptability judgment experiments: one for island effects in English WH-dependencies, one for island effects in Italian WH-dependencies, two for island effects in English RC-dependencies (two island types per experiments), and four for island effects in Italian RC-dependencies (one island type per experiment). The design of all experiments followed the factorial definition presented in Section 3.2.

There are three complications that should be noted. First, because of the potential confound of the multiple wh-word/phrase prohibition in Italian interrogatives (cf. Sec. 2.4), the interrogator island design for English WH-dependencies was constructed using the complementizer *whether* to introduce the island structure, and the corresponding Italian complementizer *se* for the corresponding island design for Italian WH-dependencies. However, the wh-island designs for both English and Italian RC-dependencies were constructed using wh-interrogative with wh-words or phrases (a mix of *when*, *where*, *how*, and *why* in English and *when*, *where*, *in which way*, *how come*, and *why* in Italian). This resulted in a regular difference between WH-dependencies, which tested *whether/se* islands, and RC-dependencies, which tested wh-islands.
Second, because we wished to introduce all of the relative clauses with relative pronouns, the Italian RC-dependencies had to be constructed from PP arguments or adjunct gaps (as opposed to the direct objects used in English RC-dependencies). In fact, Italian restrictive relative clauses (the only kind of relative clauses we used for our stimuli) only allow for the complementizer che ‘that’ with a subject or object gap. When any other argument or an adjunct is relativized, then a relative pronoun is required. Italian has two relative pronouns: the uninflected cui and il/lo/la/le/li/gl+i quale/i (which is formed by the definite determiner inflecting for number and gender and a wh-word inflecting for number only). Though etymologically related to wh-words, neither form can be used to form a wh-interrogative. Both the uninflected form (cui) and the complex form (il/lo/la/le/li/gl+i quale/i) can be used interchangeably with differences in register sometimes. In our stimuli, we chose one form or the other based on which sounded more natural to us.

Third, the subject island design in the Italian WH-dependency used a factorial definition that differs subtly from the factorial definition in (12) above. Whereas the factor GAP-POSITION in (12) varies between MATRIX-SUBJECT and EMBEDDED-OBJECT gap positions, it varies between EMBEDDED-OBJECT and EMBEDDED-SUBJECT gap positions in the Italian WH-dependency experiment. This slightly modified design closely resembles the definitions used in traditional syntax studies, but results in a non-monotonic interaction rather than a monotonic interaction. In order to standardize the interaction types (as monotonic), the subject island designs in all other experiments (English WH-dependencies and RC-dependencies, Italian RC-dependencies), use the definition in (12).

Examples of each of the conditions for each of the islands across both construction types and both languages are given in (13-28) below. English WH-dependencies are given in (13)-(16), Italian WH-dependencies in (17)-(20), English RC-dependencies in (21)-(24), and Italian RC-dependencies in (25)-(28). In all the examples below, island structures are in the c. and d. examples and are in brackets. The d. examples contain WH/RC-dependencies that cross island structures; however, we did not mark them with asterisks because the goal of this study is to experimentally determine their acceptability.

**ENGLISH: WH-dependencies**

(13) **WHETHER ISLANDS**
    a. Who __ thinks that John bought a car?
    b. What do you think that John bought __?
    c. Who __ wonders [whether John bought a car]?
    d. What do you wonder [whether John bought __ ]?

(14) **COMPLEX NP ISLANDS**
    a. Who __ heard that Jeff baked a pie?
    b. What did you hear that Jeff baked __?
    c. Who __ heard [the statement that Jeff baked a pie]?
    d. What did you [hear the statement that Jeff baked __ ]?
(15) SUBJECT ISLANDS
   a. Who __ thinks the gift prompted the congressional hearing?
   b. What do you think the gift prompted __?
   c. Who __ thinks [the gift from the lobbyist] prompted the congressional hearing?
   d. Who do you think [the gift from __] prompted the congressional hearing?

(16) ADJUNCT ISLANDS
   a. Who __ thinks that the lawyer forgot his briefcase at the office?
   b. What do you think that the lawyer forgot __ at the office?
   c. Who __ worries [if the lawyer forgets his briefcase at the office]?
   d. What do you worry [if the lawyer forgets __ at the office]?

ITALIAN: WH-dependencies

(17) WHETHER ISLANDS
   a. Chi __ pensa che io abbia letto il libro?
      who thinks that I have.SUBJ.1SG read the book
      ‘Who thinks I read the book?’
   b. Cosa pensi che io abbia letto __?
      what think.2SG that I have.SUBJ.1SG read
      ‘What do you think I read?’
   c. Chi __ si domanda [se io abbia letto il libro]?
      who to_himself asks if I have.SUBJ.1SG read the book
      ‘Who wonders if I read the book?’
   d. Cosa ti domandi [se io abbia letto __ ]?
      what to_yourself ask.2SG if I have.SUBJ.1SG read
      (‘What do you wonder if I read?’)
(18) COMPLEX NP ISLANDS

a. *Chi ha affermato che io avrei rubato una macchina?*
   ‘Who claimed that I stole a car?’

b. *Cosa hai affermato che io avrei rubato?*
   ‘What did you claim that I stole?’

c. *Chi ha fatto [l’affermazione che io avrei rubato una macchina]?*
   ‘Who made the claim that I stole a car?’

d. *Cosa hai fatto [ l’affermazione che io avrei rubato]?*
   ‘What did you make the claim that I stole?’

(19) SUBJECT ISLANDS

a. *Chi pensi che il quadro raffiguri.?*
   ‘Who do you think that the painting portrays?’

b. *Chi pensi che abbia dipinto il quadro?*
   ‘Who do you think has painted the painting?’

c. *Di chi pensi che [ il quadro sulla parete] raffiguri la nascita.?*
   ‘Who do you think the painting on the wall depicts the birth of?’

d. *Di chi pensi che [ il quadro ] raffiguri la nascita di Venere?*
   ‘Who do you think the painting depicts the birth of Venus?’
(20) **ADJUNCT ISLANDS**

a. *Chi __ dice che io abbia usato il cellulare in classe?*
   ‘Who says that I used the cell phone in class?’

b. *Cosa dici che io abbia usato __ in classe?*
   ‘What do you say that I used in class?’

c. *Chi __ si lamenta [se uso il cellulare in classe]?*
   ‘Who complains if I use my cell phone in class?’

d. *Cosa ti lamenti [se uso __ in classe]?*
   ‘What do you complain if I use in class?’

**ENGLISH: RC-dependencies**

(21) **WH-ISLANDS**

a. I take classes with the **professor who __** thinks that Paul will tutor the struggling student.

b. I take classes with the **struggling student who** the professor thinks that Paul will tutor __.

c. I take classes with the **professor who __** wonders [when Paul will tutor the struggling student].

d. I take classes with the **struggling student who** the professor wonders [when Paul will tutor __].

(22) **COMPLEX NP ISLANDS**

a. I know the **fisherman who __** heard that Laura is dating the boat captain.

b. I know the **boat captain who** the fisherman heard that Laura is dating __.

c. I know the **fisherman who __** heard [the rumor that Laura is dating the boat captain].

d. I know the **boat captain who** the fisherman heard [the rumor that Laura is dating __].
(23) **SUBJECT ISLANDS**

a. I agree with the *reporter who* __ thinks the lobbyist prompted the congressional hearing.

b. I researched the *lobbyist who* the reporter thinks __ prompted the congressional hearing.

c. I agree with the *reporter who* __ thinks [an illegal donation from the lobbyist] prompted the congressional hearing.

d. I researched the *lobbyist who* the reporter thinks [an illegal donation from __ ] prompted the congressional hearing.

(24) **ADJUNCT ISLANDS**

a. I called the *secretary who* __ thought that the lawyer insulted the client.

b. I called the *client who* the secretary thought that the lawyer insulted __.

c. I called the *secretary who* __ worries [if the lawyer insults the client].

d. I called the *client who* the secretary worries [if the lawyer insults __].

**ITALIAN: RC-dependencies**

(25) **WH-ISLANDS**

a. *Mi sono imbattuto nel ragazzo a cui* ho detto __ che myself am run into-the young_man to whom have.1sg told that hai litigato con Massimo.

have.2sg fought with Massimo

‘I ran into the young man who I told you fought with Massimo.’

b. *Mi sono imbattuto nel ragazzo con cui* penso che tu myself am run into-the young_man with whom think.1sg that you abbia litigato __.

have.sbj sg fought

‘I ran into the young man I think you fought with.’

c. *Mi sono imbattuto nel ragazzo a cui* ho domandato __ myself am run into_the young_man to whom have.1sg asked [ perché mai tu abbia litigato con Massimo].

why ever you have.sbj sg fought with massimo

‘I ran into the young man who I asked why on earth you fought with Massimo.’

d. *Mi sono imbattuto nel ragazzo con cui* mi domando myself am run into_the young_man with whom to myself ask.1sg [ perché mai tu abbia litigato __ ].

why ever you have.sbj sg fought

(‘I ran into the young man who I wonder why on earth you fought with.’)
(26) **Complex NP islands**

a. Ho telefonato alla persona alla quale hai fatto notare che Piero ha un atteggiamento ostile verso di noi.
   ‘I called the person to whom you pointed out that Piero has a hostile attitude towards us.’

b. Ho telefonato alla persona verso la quale hai fatto notare che Piero ha un atteggiamento ostile.
   ‘I called the person towards whom you pointed out that Piero has a hostile attitude.’

c. Ho telefonato alla persona alla quale hai fatto notare [il fatto che Piero ha un atteggiamento ostile verso di noi].
   ‘I called the person to whom you pointed out the fact that Piero has a hostile attitude towards us.’

d. Ho telefonato alla persona verso la quale hai fatto notare [il fatto che Piero ha un atteggiamento ostile].
   ‘I called the person towards whom you pointed out the fact that Piero has a hostile attitude.’

(27) **Subject islands**

a. Ho parlato con un collega a cui hanno riferito che hai fatto un lungo viaggio.
   ‘I spoke to a colleague who they told that you did a long trip.’

b. Ho parlato con un collega con cui mi hanno riferito che hai fatto un lungo viaggio.
   ‘I spoke to a colleague who they told me that you did a long trip with.’

c. Ho parlato con un collega a cui hanno riferito che [il tuo viaggio] è stato molto lungo.
   ‘I spoke to a colleague who they told that your trip was very long.’

d. Ho parlato con un collega con cui mi hanno riferito che [il tuo viaggio] è stato molto lungo.
   ‘I spoke to a colleague who they told that your trip was very long.’
The task in all eight experiments was magnitude estimation. We chose magnitude estimation because it has been claimed to allow participants to report more fine-grained differences in acceptability than Likert scale (e.g., 7-point) rating tasks, and is therefore compatible with the goal of this study (to uncover novel data about cross-linguistic variation). Although most of the claims about the superiority of magnitude estimation over other tasks have been empirically challenged (Bader and Haüssler 2010, Sprouse 2011b, Weskott and Fanselow 2011, Sprouse and Almeida submitted), magnitude estimation has been shown to be neither better nor worse than Likert scale tasks, therefore the choice still appears reasonable. In the magnitude estimation task participants are first presented with a reference sentence, called the standard, which is pre-assigned an acceptability rating, called the modulus. Participants are then asked to use the standard to estimate the acceptability of the experimental items. For example, if the standard is assigned a modulus of 100, and the participant believes that an experimental item is twice as acceptable as the standard, the participant would rate the experimental item as 200. If a participant believes the experimental item is half as acceptable as the standard, she would rate the experimental item as 50. The standard sentence was in the middle range of acceptability; for
example, the English standard for WH-dependencies was *Who said that my brother was kept tabs on by the FBI?*

In the remaining sections we will first report the specific experimental methods and results in four separate sections (English WH-dependencies, Italian WH-dependencies, English RC-dependencies, and Italian RC-dependencies), as we believe this is the most logical organization (Sections 4-7). We will then present a general discussion that compares the four sets of results and discusses the issues that they raise, and their implications for theories of island in both syntax and acquisition (Section 8). We will then conclude with a discussion of the implications of these results for the use of experimental syntax in cross-linguistic studies (Section 9).

### 4. English WH-dependencies

English WH-dependency island effects have been investigated several times using the factorial definition made available by experimental syntax techniques (Sprouse 2007, Sprouse 2011a, Sprouse et al. 2011, Sprouse et al. 2012). Instead of using resources to re-test a highly replicated result, here we simply report the results from Sprouse et al. 2012 (experiment 2) to establish the facts of English WH-dependency island effects.

#### 4.1 Participants

176 self-reported monolingual native speakers of English (152 Female), all University of California Irvine undergraduates, participated in this experiment for either $5 or course credit. The experiment was administered during a single visit to the lab during which the participants completed the (magnitude estimation) acceptability judgment task followed by two working memory tasks. 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.

#### 4.2 Materials

Four island types were tested: *whether*, complex NP, subject, and adjunct islands. For each island type, extraction site and structural environment was manipulated in a $2 \times 2$ design, as discussed in Section 3.2, 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 matched lexical sets of each condition were created for each island type (to ensure that there was no lexical variation, and minimal plausibility variation, across conditions). The lexically matched sets were distributed among 4 lists using a Latin Square procedure so that participants never saw lexically related items in their particular survey. 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 conditions (i.e., conditions from a $2 \times 2$ island design) 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.
4.3 Presentation

The 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. Participants were under no time constraints during their visit.

4.4 Analysis

Acceptability judgments from each participant were z-score transformed prior to further statistical analysis. The z-score transformation eliminates certain kinds of scale biases between participants (e.g., using one end of the scale, or using a larger or smaller range of values) by converting each participant’s ratings to a standardized scale (each transformed rating represents the number of standard deviations the raw rating was from the participant’s mean rating). The results were not log-transformed (contrary to some suggestions in the literature), as the z-score transformation has been shown to be both theoretically and empirically superior to the log-transformation for correcting scale bias and increasing the normality of the distribution of responses (e.g., Sprouse 2007). Differences-in-differences scores were calculated from the z-score transformed ratings.

As a statistical analysis, 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. This is 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 decided to use linear mixed effects models because of their current popularity among some experimentalists; however, it should be noted that the theoretical appropriateness of treating the items in acceptability judgment experiments as a random effect is a matter of some controversy (see Wike and Church 1976 and other articles in that volume). As such, these statistical tests may be overly conservative (i.e., the p-values reported here may be too high).

4.5 Results

In order to assess both the traditional and factorial definition of island effects, we have summarized the results of this experiment in two ways. First, Table 1 presents three pieces of information: the mean ratings of each condition of each island type, the size of the island effect as summarized by differences-in-differences scores, and the p-value of the interaction of GAP-POSITION × STRUCTURE as derived from the linear mixed effects models. The mean ratings for each condition of each island type are plotted in interaction plots in Figure 2 in order to visually confirm that any significant statistical interactions are indeed of the correct (monotonic) form, and to underscore the relative differences between conditions.
Table 1: English, WH-dependencies. Mean ratings of z-scored magnitude estimates for each condition for each island type, effect size (differences-in-differences) for each island type, and p-value of the interaction (based on linear mixed effects models) for each island type.

<table>
<thead>
<tr>
<th></th>
<th>whether</th>
<th>complex NP</th>
<th>subject</th>
<th>adjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>NON-ISLAND</td>
<td>MATRIX</td>
<td>1.23</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>NON-ISLAND</td>
<td>EMBEDDED</td>
<td>0.38</td>
<td>0.18</td>
<td>0.38</td>
</tr>
<tr>
<td>ISLAND</td>
<td>MATRIX</td>
<td>0.71</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>ISLAND</td>
<td>EMBEDDED</td>
<td>-0.73</td>
<td>-0.73</td>
<td>-0.97</td>
</tr>
<tr>
<td>DIFFERENCES-IN-DIFFERENCES</td>
<td>0.58</td>
<td>0.81</td>
<td>1.26</td>
<td>0.69</td>
</tr>
<tr>
<td>GAP-POSITION × STRUCTURE</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
<td>.0001</td>
</tr>
</tbody>
</table>

Figure 2: English, WH-dependencies. Interaction plots for each island type with p-value of the interaction term according to linear mixed effects models.
Turning first to the traditional definition of island effects, Table 1 confirms that the mean rating for the island violating condition \((\text{ISLAND} \mid \text{EMBEDDED})\) of each island type is indeed at or below -0.5 (the criterion established in Section 3 in order to ensure that the unacceptability of traditionally defined islands would be visible without the need for sensitive statistical tests). This corroborates previous results in the syntax literature that used traditional methods to define island effects for English WH-dependencies. Turning next to the factorial definition, Table 1 confirms that every island type leads to a significant interaction (at \(p < .0001\)) using linear mixed effects models, and that the size of each interaction is at or above 0.5 as calculated by DD scores (the criterion established in Section 3.2 in order to ensure that the interaction would be visible without the need for sensitive statistical tests). The interaction plots in Figure 2 also confirm that each of the interactions are of the correct form: super-additive monotonic interactions that indicate that the unacceptability of the island violating sentence \((\text{ISLAND} \mid \text{EMBEDDED})\) of each island type is greater than would be predicted based on the simple linear summation of the acceptability costs of the two factors alone. In short, there are clear island effects for English WH-dependencies under both definitions for all four island types.

5. Italian WH-dependencies

5.1 Participants

144 self-reported native speakers of Italian participated in the experiment. They were all undergraduate students at the University of Milano-Bicocca in Milan, Italy. Because Italian educational policies prevent payment of students for participation in research experiments, all participants were volunteers. Three participants were excluded from the analysis for obviously performing the task incorrectly.

5.2 Materials

The same four island types were tested in this experiment. For \textit{whether/se}, complex NP, and adjunct islands, extraction site and structural environment was manipulated in a 2\(\times\)2 design, as discussed in Section 3.2. Subject islands were tested in a slightly different design: the gap-location was manipulated between subject and object position of the embedded clause (rather than between the subject of the matrix clause and the subject of the embedded clause). This design more closely resembles the typical subject-island paradigm in the syntactic literature; however, it also leads to a non-monotonic interaction causing a slightly different interaction graph. For each island type, 16 lexically matched sentence sets were created and distributed among 8 lists using a Latin square design. This ensured that each participant rated 2 tokens of each condition, but never saw related lexicalizations within or across conditions. The 32 target items (4 islands \(\times\) 4 conditions \(\times\) 2 token) in each list were combined with 28 unrelated filler items (which served as experimental conditions for a separate experiment). The filler items were chosen such that the composition of each survey was 50% acceptable and 50% unacceptable (by hypothesis). Three novel orders were created for each list by pseudorandomizing the 60 items such that related conditions (i.e., conditions from the same island type) never appeared consecutively. This resulted in 24 different surveys. The standard, given in (29) below, was identical for all 24 surveys, and was in the middle range of acceptability. The standard was assigned a modulus of 100.
(29) Cosa hai chiesto se tua madre ha comprato per tuo padre?  
what have.2sg asked if your mother has bought for your father  
(‘What did you ask if your mother bought for your father?’)

5.3 Presentation

The 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 3 items of the survey were practice items. 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 63 items long. Participants completed the surveys during a class period, but were otherwise under no time constraints.

5.4 Analysis

Ratings from each participant were z-score transformed prior to statistical analysis in order to eliminate certain kinds of scale bias. The z-score transformed ratings were then analyzed using linear mixed effects models with items and participants included as random factors for each island type using GAP-POSITION and STRUCTURE as fixed factors. All p-values were estimated using the MCMC method implemented in the languageR package for R (Baayen 2007, Baayen et al 2008).

5.5 Results

In order to assess both the traditional and factorial definition of island effects, we have summarized the results of this experiment in two ways. First, Table 2 presents three pieces of information: the mean ratings of each condition of each island type, the size of the island effect as summarized by differences-in-differences scores, and the p-value of the interaction of GAP-POSITION × STRUCTURE as derived from the linear mixed effects models. The mean ratings for each condition of each island type are plotted in interaction plots in Figure 3 in order to visually confirm that any significant statistical interactions are indeed of the correct (monotonic) form, and to underscore the relative differences between conditions.
Table 2: Italian, WH-dependencies. Mean ratings of z-scored magnitude estimates for each condition for each island type, effect size (differences-in-differences) for each island type, and p-value of the interaction (based on linear mixed effects models) for each island type. The design for subject islands was slightly different. The level names for the GAP-POSITION factor of the subject island design are given in parentheses.

<table>
<thead>
<tr>
<th>Condition Type</th>
<th>whether</th>
<th>complex NP</th>
<th>subject</th>
<th>adjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>NON-ISLAND</td>
<td>MATRIX (or OBJECT)</td>
<td>0.83</td>
<td>0.39</td>
<td>-0.17</td>
</tr>
<tr>
<td>NON-ISLAND</td>
<td>EMBEDDED (or SUBJECT)</td>
<td>0.97</td>
<td>-0.43</td>
<td>1.04</td>
</tr>
<tr>
<td>ISLAND</td>
<td>MATRIX (or OBJECT)</td>
<td>0.08</td>
<td>0.42</td>
<td>-0.45</td>
</tr>
<tr>
<td>ISLAND</td>
<td>EMBEDDED (or SUBJECT)</td>
<td>-1.00</td>
<td>-0.97</td>
<td>-0.83</td>
</tr>
<tr>
<td>DIFFERENCES-IN-DIFFERENCES</td>
<td></td>
<td>1.22</td>
<td>0.51</td>
<td>1.59</td>
</tr>
<tr>
<td>GAP-POSITION × STRUCTURE</td>
<td></td>
<td>.0001</td>
<td>.0006</td>
<td>.0001</td>
</tr>
</tbody>
</table>

Figure 3: Italian, WH-dependencies. Interaction plots for each island type with p-value of the interaction term according to linear mixed effects models.
Turning first to the traditional definition of island effects, Table 2 confirms that the mean rating for the island violating condition (ISLAND | EMBEDDED or SUBJECT) of each island type is indeed at or below -0.5 (the criterion established in Section 3 in order to ensure that the unacceptability of traditionally defined islands would be visible without the need for sensitive statistical tests). This corroborates previous results in the syntax literature that used traditional methods to define island effects for Italian WH-dependencies. Turning next to the factorial definition, Table 2 confirms that every island type leads to a significant interaction (at \( p < .0001 \) for all but complex NP islands, which were \( p < .0006 \)) using linear mixed effects models, and that the size of each interaction is at or above 0.5 as calculated by DD scores (the criterion established in Section 3.2 in order to ensure that the interaction would be visible without the need for sensitive statistical tests). The interaction plots in Figure 3 also confirm that each of the interactions are of the correct form: super-additive monotonic interactions that indicate that the acceptability of the island violating sentence (ISLAND | EMBEDDED) is lower than would be predicted based on the simple linear summation of the acceptability costs of the two factors alone for whether, complex NP, and adjunct islands, and a non-monotonic interaction that indicates that the (ISLAND | SUBJECT) condition is less acceptable than would be predicted. In short, there are clear island effects for Italian WH-dependencies under both definitions for all four island types.

Although the focus of this article is on island effects, it should be noted that two of the purportedly acceptable conditions in this experiment were rated relatively close to the unacceptable criterion of -0.5: the NON-ISLAND | EMBEDDED condition of the complex NP island design (-0.43), and the ISLAND | OBJECT condition of the subject island design (-0.45). The first fact suggests that in Italian there may be something troubling about extraction from an embedded clause that is the complement of verbs like *mettere in dubbio* ‘to doubt’ (lit. ‘to put into doubt’), *insinuare* ‘insinuate’, *raccontare* ‘to tell’, *negare* ‘deny’, *affermare* ‘to claim’, *dubitare* ‘to doubt’, *richiedere* ‘to request’, and *sentire* ‘to hear’. The second fact may suggest independent difficulties in extracting PPs within NPs in Italian, regardless of the structural position of the NP in the clause. We can leave these as questions for future research because the factorial definition of island effects is explicitly designed to control for these types of differences in acceptability: a significant interaction (of the appropriate type and size) suggests that island violating sentences are less acceptable than would be predicted even after accounting for these relatively unacceptable sentence types.

6. English RC-dependencies

Island effects with English RC-dependencies were tested in two experiments: one experiment testing wh and subject islands, and one experiment testing complex NP and adjunct islands. The rationale for the division was that relative clauses may be more salient than wh-interrogatives to participants in a judgment study, and we didn’t want to risk fatigue or boredom that may lead to less accurate ratings.
6.1 Participants

64 self-reported native speakers of English participated in each experiment (128 participants total) through the Amazon Mechanical Turk website. Participants were paid $2 for their participation. Participant selection criteria were enforced as follows. First, the AMT interface automatically restricted participation to AMT users with a US-based location. Second, we included two questions at the beginning of the experiment to assess language history: (1) Were you born and raised in the US?, (2) Did both of your parents speak English to you at home? These questions were not used to determine eligibility for payment, consequently there was no incentive to lie. 1 participant was removed from the wh/subject experiment, and 3 participants were removed from the complex np/adjunct experiment, for either answering ‘no’ to one or both of these questions or for obvious attempts to receive payment without honestly performing the task.

6.2 Materials

For each island type, extraction site and structural environment was manipulated in a 2×2 design, as discussed in Section 3.2, yielding a total of 8 critical conditions in each experiment. 8 matched lexical sets of each condition were created for each island type (to ensure that there was no lexical variation, and minimal plausibility variation, across conditions). The lexically matched sets were distributed among 8 lists using a Latin Square procedure so that participants never saw lexically related items in their particular survey. This meant that each list consisted of 1 token per sentence type. The same 24 additional filler sentences were added to each list to both add variety to the surveys, and to ensure that the full range of acceptability judgments were likely to be used. The distribution of the filler items across the range of acceptability was balanced to ensure an equal number of acceptable and unacceptable sentences. The order of sentences within each list was pseudorandomized such that related conditions (i.e., conditions from a 2×2 island design) 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.

6.3 Presentation

The task was presented as an online survey on the Amazon Mechanical Turk website. The experiment began with a practice phase during which participants estimated the lengths of 6 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, 7 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. 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 41 items long. Participants completed the surveys using a web browser at their own pace.
6.4 Analysis

Ratings from each participant were z-score transformed prior to statistical analysis in order to eliminate certain kinds of scale bias. The z-score transformed ratings were then analyzed using linear mixed effects models with items and participants included as random factors for each island type using GAP-POSITION and STRUCTURE as fixed factors. All $p$-values were estimated using the MCMC method implemented in the languageR package for R (Baayen 2007, Baayen et al 2008).

6.5 Results

In order to assess both the traditional and factorial definition of island effects, we have summarized the results of this experiment in two ways. First, Table 3 presents three pieces of information: the mean ratings of each condition of each island type, the size of the island effect as summarized by differences-in-differences scores, and the $p$-value of the interaction of GAP-POSITION $\times$ STRUCTURE as derived from the linear mixed effects models. The mean ratings for each condition of each island type are plotted in interaction plots in Figure 4 in order to visually confirm that any significant statistical interactions are indeed of the correct (monotonic) form, and to underscore the relative differences between conditions.

**Table 3: English, RC-dependencies.** Mean ratings of z-scored magnitude estimates for each condition for each island type, effect size (differences-in-differences) for each island type, and $p$-value of the interaction (based on linear mixed effects models) for each island type.

<table>
<thead>
<tr>
<th></th>
<th>wh</th>
<th>complex NP</th>
<th>subject</th>
<th>adjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>NON-ISLAND</td>
<td>MATRIX</td>
<td>0.34</td>
<td>0.34</td>
<td>0.52</td>
</tr>
<tr>
<td>NON-ISLAND</td>
<td>EMBEDDED</td>
<td>-0.17</td>
<td>-0.38</td>
<td>0.13</td>
</tr>
<tr>
<td>ISLAND</td>
<td>MATRIX</td>
<td>-0.20</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>ISLAND</td>
<td>EMBEDDED</td>
<td>-0.70</td>
<td>-0.64</td>
<td>-0.68</td>
</tr>
<tr>
<td>DIFFERENCES-IN-DIFFERENCES</td>
<td>-0.01</td>
<td>0.15</td>
<td>0.73</td>
<td>0.30</td>
</tr>
<tr>
<td>GAP-POSITION $\times$ STRUCTURE</td>
<td>.926</td>
<td>.042</td>
<td>.0068</td>
<td>.1018</td>
</tr>
</tbody>
</table>
Turning first to the traditional definition of island effects, Table 3 indicates that the mean rating for the island violating condition (ISLAND | EMBEDDED) of each island type is indeed at or below -0.5 (the criterion established in Section 3 in order to ensure that the unacceptability of traditionally defined islands would be visible without the need for sensitive statistical tests). This corroborates previous results in the syntax literature that used traditional methods to define island effects for English RC-dependencies.

Turning next to the factorial definition of island effects, Table 3 and Figure 4 reveal a different set of results. Whereas subject islands yield a clear island effect under the factorial definition (i.e., a significant interaction \( p < .007 \) that is of the predicted form and size (DD is 0.73)), the other three island types do not. Wh-islands yield no interaction of any sort \( (p < .93 \) and DD is nearly zero). Complex NP islands yield a significant interaction \( (p < .042 \), but it is a non-monotonic interaction in which the ISLAND | MATRIX condition is more acceptable than the NON-ISLAND | MATRIX condition. Furthermore the hypothetically acceptable NON-ISLAND | EMBEDDED condition is rated relatively unacceptable (-0.38), resulting in an interaction size that does not meet our criterion (DD is 0.15). Finally, although Figure 4 suggests an interaction of the
correct form for adjunct islands, the result of the linear mixed effects model is not significant \( p < .102 \), and the size of the interaction is smaller than our criterion (DD is 0.30). In other words, based on the factorial definition, only subject islands potentially require a grammatical constraint to explain their unacceptability under RC-dependencies; wh-islands and adjunct islands can potentially be explained by the simple linear summation of the acceptability costs of the two factors, and complex NP islands require additional investigation.

7. **Italian RC-dependencies**

Island effects with Italian RC-dependencies were tested in four experiments: one experiment for each island type. The rationale for the division was that relative clauses may be more salient than wh-interrogatives to participants in a judgment study, and we didn’t want to risk fatigue or boredom that may lead to less accurate ratings.

7.1 **Participants**

69 self-reported native speakers of Italian participated in the *whether* island experiment, 68 in the complex NP island experiment, 53 in the subject island experiment, and 52 in the adjunct island experiment. They were all undergraduate students at University of Milano-Bicocca in Milan, Italy. Because Italian educational policies prevent payment of students for participation in research experiments, all participants were volunteers. No participants were excluded from the analysis.

7.2 **Materials**

For each island type/experiment, extraction site and structural environment was manipulated in a \( 2 \times 2 \) design, as discussed in Section 3.2, yielding a total of 4 critical conditions in each experiment. 8 matched lexical sets of each condition were created for each island type (to ensure that there was no lexical variation, and minimal plausibility variation, across conditions). The lexically matched sets were distributed among 8 lists using a Latin Square procedure so that participants never saw lexically related items in their particular survey. This meant that each list consisted of 1 token per sentence type. The four critical items in each list were combined with 16 conditions from unrelated experiments and 20 additional unrelated filler items. The unrelated conditions and filler items were chosen such that the composition of each survey was 50% acceptable and 50% unacceptable (by hypothesis), as well as 50% declarative and 50% interrogative. The filler items were identical for each experiment (and each list within the experiment), so that the filler items contributed the same variation to every participant. The order of sentences within each list was pseudorandomized so that related conditions (i.e., conditions from a \( 2 \times 2 \) island design) were never presented successively. This resulted in 8 different surveys. The standard was the same as the one given in (29) for the WH-dependencies. It was identical for all 8 surveys, and was in the middle range of acceptability. The standard was assigned a modulus of 100.
7.3 Presentation

The task was presented as a paper survey. The experiment began with a practice phase during which participants estimated the lengths of 6 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. 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 49 items long. Participants completed the surveys during a class period, but were otherwise under no time constraints.

7.4 Analysis

Ratings from each participant were z-score transformed prior to statistical analysis in order to eliminate certain kinds of scale bias. The z-score transformed ratings were then analyzed using linear mixed effects models with items and participants included as random factors for each island type using GAP-POSITION and STRUCTURE as fixed factors. All *p*-values were estimated using the MCMC method implemented in the languageR package for R (Baayen 2007, Baayen et al 2008).

7.5 Results

In order to assess both the traditional and factorial definition of island effects, we have summarized the results of this experiment in two ways. First, Table 4 presents three pieces of information: the mean ratings of each condition of each island type, the size of the island effect as summarized by differences-in-differences scores, and the *p*-value of the interaction of GAP-POSITION × STRUCTURE as derived from the linear mixed effects models. The mean ratings for each condition of each island type are plotted in interaction plots in Figure 5 in order to visually confirm that any significant statistical interactions are indeed of the correct (monotonic) form, and to underscore the relative differences between conditions.

| Table 4: Italian, RC-dependencies. Mean ratings of z-scored magnitude estimates for each condition for each island type, effect size (differences-in-differences) for each island type, and *p*-value of the interaction (based on linear mixed effects models) for each island type. |
|---------------------------------|---------|----------|----------|---------|
| NON-ISLAND | MATRIX               | wh      | complex NP | subject  | adjunct |
| NON-ISLAND | EMBEDDED             | 0.67    | 0.69      | 0.53     | 0.52    |
| ISLAND   | MATRIX               | 0.20    | 0.44      | -0.36    | -0.52   |
| ISLAND   | EMBEDDED             | 0.54    | 0.86      | 0.53     | 0.53    |
| DIFFERENCES-IN-DIFFERENCES     | 0.43    | 0.51     | -0.13     | 0.74    |
| GAP-POSITION × STRUCTURE        | .0546   | .0308    | .0094     | .0022   |
Figure 5: Italian, RC-dependencies. Interaction plots for each island type with $p$-value of the interaction term according to linear mixed effects models.

Turning first to the traditional definition of island effects, Table 4 indicates that the mean rating for complex NP and adjunct islands are at or below our pre-established criterion (-0.52 and -0.51 respectively), but subject islands are not (-0.18), and wh-islands are marginal at best (-0.36). This corroborates previous results in the syntax literature that used traditional methods to define island effects for Italian RC-dependencies (Rizzi 1982).

Turning next to the factorial definition of island effects, Table 4 and Figure 5 appear to corroborate these findings. Complex NP and adjunct islands both yield significant interactions ($p < .031, p < .0022$) of the correct form and size (DD is 0.51 and 0.74 respectively). Subject islands demonstrate a significant interaction ($p < .01$); however, that interaction is not only of the wrong form (non-monotonic), but it goes in the wrong direction (resulting in a negative DD score of -0.13), which means the majority of the difference between conditions doesn’t even involve the putative island violating condition. Finally, wh-islands demonstrate a marginal interaction ($p < .055$) that is also marginal in effect size (DD is 0.43). In the case of Italian RC-
dependencies, both the traditional and factorial definitions yield similar results: complex NP and adjunct island effects exist, subject island effects do not, and wh-island effects likely require additional investigation because they are marginal by our pre-established criteria.

8. General Discussion

In this study we looked for experimental evidence of island effects in two languages (English and Italian), across two types of dependencies (WH-dependencies and RC-dependencies), and for four types of islands (whether/wh, complex NP, subject, and adjunct). We also compared two definitions of island effects: the traditional definition, which is generally used in informal judgment experiments and involves making a judgment about the absolute acceptability of the island violating sentences, and the factorial definition, which is licensed by formal rating experiments and allows us to control for the relative acceptability of the syntactic factors that make up island effects. In order to translate the traditional definition into experimental terms, we set the criterion for unacceptability at a rating of at least -0.5 on a z-score transformed scale, as this roughly coincides with a difference between marginal acceptability (a rating of 0 on a z-score transformed scale) that would be visible without the need for statistical tests (Cohen 1988). Similarly, we defined positive evidence for island effects under the factorial definition as a significant interaction of GAP-POSITION × STRUCTURE and a differences-in-differences (effect size) of at least 0.5 on a z-score transformed scale. An effect size criterion was included because statistical significance can be achieved for very small effects by simply testing a large enough sample of participants, and the definition of island effects in the syntactic literature suggests that they are not small effects (i.e., there is a difference between statistical significance in a test and practical significance for a theory, as already discussed in Section 3.2). Once again, the effect size criterion was chosen to coincide with an effect that is large enough to be visible without statistics.

8.1 The pattern of results

Turning first to the pattern of results returned by the traditional definition in Table 5, we can see that these experiments straightforwardly corroborated the results that have been reported in the syntactic literature using traditional (informal) judgment collection techniques. Both English and Italian WH-dependencies demonstrate all four island types, English RC-dependencies demonstrate all four island types, and Italian RC-dependencies only demonstrate complex NP and adjunct islands, not wh- and subject islands (wh-islands are marginal, which we will interpret as non-significant here so as to be as conservative as possible).

Table 5: Qualitative results for the traditional definition of island effects. The presence of an island effect is indicated by a plus (+), the absence by a minus (-), and the marginal presence by the word marginal.

<table>
<thead>
<tr>
<th>language</th>
<th>dependency</th>
<th>whether/wh</th>
<th>complex NP</th>
<th>subject</th>
<th>adjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>WH-dependencies</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Italian</td>
<td>WH-dependencies</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>English</td>
<td>RC-dependencies</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Italian</td>
<td>RC-dependencies</td>
<td>marginal</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
The agreement between these results and previous results is not surprising under the view that traditionally collected judgments are a reliable form of informal experiment (e.g., Marantz 2005, Phillips 2009, Sprouse and Almeida in press a, Sprouse, Schütze, and Almeida submitted, Sprouse and Almeida in press b). And if there were no other possible definitions of island effects, this would be strong corroboration of the constrained variation that has been reported in the syntactic literature and captured by grammatical mechanisms like parameterized bounding-nodes and the Subjacency Condition (Rizzi 1982, Torrego 1984, Chomsky 1986).

Turning next to the pattern of results returned by the factorial definition in Table 6, we see an interesting combination of similarities and differences between the traditional and factorial definitions.

Table 6: Qualitative results for the factorial definition of island effects. The presence of an island effect is indicated by a plus (+), the absence by a minus (-), and the marginal presence by the word marginal.

<table>
<thead>
<tr>
<th>language</th>
<th>dependency</th>
<th>whether/wh</th>
<th>complex NP</th>
<th>subject</th>
<th>adjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>WH-dependencies</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Italian</td>
<td>WH-dependencies</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>English</td>
<td>RC-dependencies</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>marginal</td>
</tr>
<tr>
<td>Italian</td>
<td>RC-dependencies</td>
<td>marginal</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

The factorial definition returned the same results as the traditional definition for English WH-dependencies, Italian WH-dependencies, and Italian RC-dependencies: all four island types are present for English and Italian WH-dependencies, and only complex NP and adjunct islands are present for Italian RC-dependencies. However, the factorial definition returned different results than the traditional definition for English RC-dependencies. Under the factorial definition, only subject islands are clearly present in English RC-dependencies, while wh, complex NP, and adjunct islands appear to be absent (adjunct islands are marginal, but we will interpret that as absent in order to be as conservative as possible). A comparison of the results of the two definitions is given in Table 7.

Table 7: Comparison of the qualitative results for the traditional and the factorial definitions of island effects. The presence of an island effect is indicated by a plus (+), the absence by a minus (-), and the marginal presence by the word marginal. If the results for the traditional and the factorial definitions don’t differ, only one symbol per cell is reported. If they do differ, “traditional” results are reported to the left of “/” and the corresponding “factorial” results are reported to the right of “/”, and they are both highlighted in bold.

<table>
<thead>
<tr>
<th>language</th>
<th>dependency</th>
<th>whether/wh</th>
<th>complex NP</th>
<th>subject</th>
<th>adjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>WH-dependencies</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Italian</td>
<td>WH-dependencies</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>English</td>
<td>RC-dependencies</td>
<td>+/-</td>
<td>+/-</td>
<td>+</td>
<td>+/-marginal</td>
</tr>
<tr>
<td>Italian</td>
<td>RC-dependencies</td>
<td>marginal</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
8.2 Interpreting the two definitions in tandem

Although on the surface it appears as the two definitions are in conflict with respect to English RC-dependencies, the factorial definition in fact provides more information than the traditional definition. This means that the factorial definition does not conflict with the traditional definition, but rather refines it. In this case, the factorial definition indicates that the low acceptability associated with wh-, complex NP, and adjunct island violations in English RC-dependencies can potentially be explained by the simple linear summation of the acceptability costs associated with the structural factors that make up island violations: a long-distance dependency and a specific type of syntactic structure. For wh- and adjunct islands, each of these factors has a cost, and combining them accurately predicts the acceptability of island violations. For complex NP islands, it appears as though there is no cost for the syntactic structure; instead, the cost for long distance dependencies accounts for nearly all of the unacceptability of the island violation.

Because the unacceptability of these “island violations” can be explained by the sum of the (one or) two structural factors that define them, there is no reason to invoke additional grammatical constraints to explain the low acceptability. In other words, as far as English RC-dependencies are concerned, only subject islands appear to require an extra grammatical constraint to explain the low acceptability returned by the traditional definition; the low acceptability of the other three islands can be explained without extra grammatical constraints.

8.3 Consequences for syntactic theories

The results revealed by the factorial definition of island effects raise potential problems for previous theories of variation in island effects. For example, the classic Subjacency approach (Rizzi 1982, Torrego 1984) of parameterized bounding node options was designed to capture covariation between wh- and subject islands: if CP were a bounding a node, then the language would lack both wh- and subject islands, if IP were a bounding node, then the language would demonstrate both wh- and subject islands. The results of these experiments suggest that it is empirically possible to lack wh-islands, but still demonstrate subject islands. Although we do not yet know whether the reverse is possible (lack subject islands, but demonstrate wh islands), the existence of this pattern is a problem for any approach that posits a direct link (and thus direct covariation) between wh- and subject islands. We should also note that these results are compatible with an approach that posits a direct link between complex NP and adjunct islands. However, there is a clear asymmetry in the value of this pattern as evidence for a direct link, and the previous pattern as evidence against a direct link. The previous pattern (the presence of subject islands and the absence of wh islands) is a clear counterexample to any proposal that postulates direct covariation. As long as they can’t be explained away, counterexamples are strong evidence against a claim. The strength of evidence in support of claim, such as the presence of two islands (or the absence of two islands), is only as strong as the effort one expends in searching for counterexamples. In this case, we only investigated two languages, which isn’t a very comprehensive search for a counterexample to the claim that complex NP and adjunct islands co-vary.

The results revealed by the factorial definition also raise potential problems for a more abstract claim in syntactic theories: that WH-dependencies and RC-dependencies are two instances of a single phenomenon, at least with respect to island constraints. We know of no previous study that has argued that WH-dependencies and RC-dependencies in a single language
should be subject to different island effects. The Italian data presented by Rizzi (1982) would have been an example of this, but rather than endorse such a claim, Rizzi argued that the wh-island effects in Italian WH-dependencies were actually the result of a prohibition on multiple-wh constructions in Italian. Rizzi’s explanation for the variation between WH-dependencies and RC-dependencies is difficult to maintain for the other island types; however, we no longer need to look at Italian as an example of this sort of variation. Under the factorial definition of island effects, English shows clear variation between WH-dependencies and RC-dependencies. English has no general prohibition on multiple-wh constructions, so Rizzi’s (1982) explanation is not possible. The only conclusion is that WH-dependencies and RC-dependencies can vary with respect to island effects within a single language. This, of course, does not entail that they are completely separate phenomena. For example, it is logically and empirically possible that they are both generated by the same syntactic mechanism (movement, copy and internal merge, slash passing, etc.), but that different subsets of constraints apply to each construction type. This variation has potential ramifications for the theory of acquisition, which we turn to in section 8.5.

8.4 Consequences for the debate between grammatical approaches and reductionist approaches to island effects

These results also have potential consequences for the debate between grammatical approaches and reductionist approaches to the source of island effects. The fact that the low acceptability of wh-, complex NP, and adjunct “island violations” in English RC-dependencies can potentially be explained by the simple linear summation of the structural factors that compose island effects (i.e., there was no superadditive interaction) corroborates the plausibility of reductionist approaches. At the same time, the superadditive interaction observed with subject islands in English RC-dependencies, and with all four island types with English WH-dependencies, corroborates the plausibility of grammatical approaches. Taken together, the existence of both superadditive and linearly additive effects potentially suggests that both grammatical and reductionist acceptability effects exist. To our knowledge, no syntactician has ever claimed that all acceptability effects are due to grammatical constraints, so this is likely not problematic for many syntacticians. However, there have been some that have claimed that the superadditive interaction observed with English WH-dependencies can be reduced to properties of the behavior of the human sentence parser without the need for grammatical constraints. While this claim is still logically possible in the face of these results, the fact that both patterns arise in English, and the fact that these researchers would likely claim that both patterns are generated without grammatical constraints, means that the presence or absence of an interaction is not evidence for or against the reductionist hypothesis. In this way, the current results either argue for the existence of grammatical constraints for some islands (but not necessarily all), or these results argue that evidence for or against the reductionist hypothesis cannot come from the superadditive pattern of acceptability seen here, but rather from other types of evidence (e.g., the working memory evidence presented by Sprouse et al. 2012 that suggests that reductionist approaches cannot rely on working memory as an explanation for island effects).

In a similar vein, the fact that these results raise problems for syntactic theories that predict constrained variation means that constrained variation may no longer be an argument in favor of grammatical approaches over reductionist approaches. Prima facie, this appears to be a positive development for reductionist approaches. However, this experiment did reveal variation both between the two languages, and within each language (even setting aside Italian WH-
dependencies because of Rizzi’s 1982 concerns about multiple-wh constructions). English WH-dependencies demonstrate different island effects than English RC-dependencies, and English RC-dependencies demonstrate different island effects than Italian RC-dependencies. This type of variation is relatively straightforward to capture under a grammatical approach to island effects, but is not easy to capture under a reductionist approach to island effects. The core explanatory power of reductionist approaches lies in their invocation of cognitive mechanisms that are independent of the phenomenon of interest. As we have seen, many prominent reductionist approaches to island effects invoke constraints on the capacity of working memory in an attempt to explain island effects without postulating any cognitive mechanisms that are specific to island effects. However, variation that appears to only occur with island effects is difficult to rectify with a general cognitive mechanism like working memory, as it is unlikely that speakers of different languages systematically vary in their working memory capacity, and it is unlikely that the working memory capacity of speakers of English varies depending on which type of construction they are speaking. The theoretical burden on reductionist theories in the face of this kind of variation between and within languages is to find a way to explain the variation without postulating grammatical mechanisms.

8.5 Consequences for theories of acquisition

Constrained syntactic variation across languages suggests that the learning strategy employed by children is constrained in such a way that it only gives rise to the observed pattern, as the alternative is the assumption that all of the languages of the world coincidentally converged on a set of input that gave rise to the constrained pattern of variation. The claim in the syntactic literature that English and Italian illustrated a type of constrained variation (direct covariation between wh- and subject islands) was therefore a piece of evidence in favor of a learning strategy that is constrained in some way with respect to islands. The fact that these experiments do not reveal that pattern of variation casts doubt on that conclusion: the presence or absence of specific island effects in each language may be determined by the input that children receive (for a learning strategy that works this way, see Pearl and Sprouse to appear). Furthermore, the fact that these experiments reveal variation across constructions in both languages suggests that WH-dependencies and RC-dependencies are treated differently by the acquisition process. This need not entail entirely distinct learning strategies, as it is possible that island effects for both constructions derive from the same strategy, but that the strategy is applied to distinct sets of input.

9. Conclusion

Cross-linguistic variation is critical to (at least) two debates at the forefront of linguistic theory. First, to what extent can complex syntactic phenomena be explained with domain-general cognitive mechanisms rather than specific syntactic mechanisms? Second, to what extent can the acquisition of complex syntactic phenomena be achieved with domain-general learning strategies rather than specific syntactic strategies? Our goal in this article was to apply formal acceptability judgment experiments (experimental syntax) to a high-profile example of cross-linguistic variation (syntactic island effects in English and Italian) in order to demonstrate both the benefits and limits of these techniques. Crucially, we compared both the traditional definition of island effects and the factorial definition made available by formal experiments to both
WH-dependencies and RC-dependencies in English and Italian in an attempt to reveal both a finer grained pattern of island effects across languages and dependency types, as well as provide additional information about the potential role of grammatical constraints in these patterns. The results suggest a previously unobserved pattern of results for English RC-dependencies, where the only genuine islands turn out to be subject islands, while for English WH-dependencies all four environments we tested behave like islands. The results also suggest a previously unobserved pattern in Italian: subjects islands are present for WH-dependencies, but not for RC-dependencies. These findings have consequences both for debates about the role of domain-general mechanisms in the existence of island effects, and for debates about the mechanisms underlying language acquisition. We also take our findings to strongly indicate that formal acceptability judgment experiments, and crucially the factorial definitions that they enable, can play a substantive role in future cross-linguistic studies.

Although these results suggest much to be optimistic about, they also highlight many of the logical limitations of experiments, both formal and informal. For example, whereas factorial definitions can indicate whether a grammatical constraint is potentially required (through an interaction), or whether reductionist approaches are potentially possible (through the lack of an interaction), no experimental result can provide the content of the analysis. As is well-known, grammatical constraints can take any number of forms, from syntactic constraints to pragmatic constraints, or from constraints on dependencies to Rizzi’s (1982) proposed constraint on the number of wh-words in the same clause. Similarly, reductionist approaches can invoke any number of domain-general cognitive mechanisms to explain a given acceptability pattern. These results also highlight the fact that statistical significance alone is never enough to interpret the theoretical implications of a result. The researcher must also evaluate the practical significance of the results. In these results, both the pattern of the results (i.e., the factorial definition) and the size of the effect were used to interpret the practical significance of the results. Finally, these results highlight the fact that complete answers are rarely available immediately. Even with a series of four experiments covering two languages, two constructions, and four island types, we still encountered two marginal results and two unpredicted acceptability effects that will require future research to fully understand.
References


Gieselman, Simone, Robert Kluender, and Ivano Caponigro. to appear. Isolating processing factors in negative island contexts. *Proceedings of NELS 41*


Sprouse, Jon and Diogo Almeida. submitted. Power in acceptability judgment experiments and the reliability of data in syntax.


