The effect of filler complexity across main- and embedded clauses on the acceptability of *wh*-island extraction in English.

Julian Rutjes

Abstract

The acceptability of syntactic island violations is generally taken to ameliorate when wh-filler complexity is increased. This effect is exclusively found in weak islands. English, for example, has weak islands, because it has filler complexity amelioration. Beljon et al. (2021) posit that Dutch has strong islands, because they found no convincing evidence for a filler complexity effect in Dutch wh-islands. In their analysis, they conclude that filler-complexity amelioration effects do not appear in Dutch main clauses, because a featural Relativized Minimality analysis shows that the Dutch grammar does not distinguish between complex or bare fillers in main clauses. The Dutch main clause verb-second (V2) constraint is suggested as the reason, because V2 ostensibly causes the C-position in Dutch main clauses to always carry some extra feature F, in addition to a Q feature. When the filler takes this position, it has full featural overlap with the embedded subject, which entails the highest degree of intervention effect, meaning strong islands. This analysis makes the clear prediction that sentences without V2 constraints (e.g. Dutch embedded clauses) should lack the feature that V2 adds to the C-position. Then, Inclusion can be avoided, leading to a weaker intervention effect, and consequently, weaker islands. Theoretically, this suggests that Dutch embedded clauses could show complexity-amelioration. The other side of this prediction is that languages without V2, such as English must be shown to have a complexity effect that stays consistent in both main and embedded clauses, because otherwise the factor V2 cannot be singled out as the cause for strong islands in Dutch. The consistency of the complexity effect between main and embedded clauses in English is what the present study aimed to empirically test. To ascertain this, a graded acceptability judgment task was conducted. The effect of complexity was found to be significant and consistent across both types of clauses, which aligns with the predictions made by the analysis in Beljon et al. that non-V2 island extractions should show consistent complexity effects across both main and embedded clauses.

Keywords: syntactic islands, wh-movement, filler complexity, Relativized Minimality embedded questions, graded judgments

1. Introduction

Syntactic movement is a widely accepted phenomenon in generative linguistics where constituents of a sentence are interpreted in a different position from where they were first merged. A prominent example of syntactic movement can be found in question sentences. Question sentences contain an element that causes it to be interpreted as a question. In linguistics, we call this element a *wh*-element (a clever name, considering that most of these elements start with *wh* – take *who, what,* or *where* for instance). In question sentences, the *wh*-element originates as the object of the verb, but is moved to the front of the sentence. See example (1).

(1) a. What did Mary eat _?
b. What does John think [Mary ate _?]

In the sentences in example (1), the *wh*-element *what* originates from the position of the underscore, where it is the object of the verb *to eat*. It subsequently moves to the front of the sentence, where it resides when the sentence is interpreted. The example sentence in (1b) shows that the *wh*-element can even move from an embedded clause to the main clause. The story thus far is largely unproblematic, but things quickly take a turn for the worse when a second *wh*-element is introduced.

Multiple-*wh* questions are questions that ask multiple things, by virtue of containing more than one *wh*-element. In many languages, the acceptability of multiple *wh*-questions degrades, when a *wh*-element has to move across another *wh*-element. In such situations, the element that has to be crossed over is referred to as an intervening element or intervener. This process is illustrated in (2). In both examples, *what* moves to the front of the sentence, even though it was originally generated at the object position of *eat*(2a)*/ate*(2b). When it moves to the front of the sentence, it has to move past the intervening *wh*-element *who* to be fronted. As a result of this movement, native speakers generally grade (2a) to be questionable, but still interpretable, while (2b) is considered unacceptable by any account.

(2) a. ??What did who eat _?
b. *What do you wonder [who ate _?]

An astute observer might question why (2b) is graded so much worse than (2a). The key difference can be found in the structure of the two sentences. In (2b), the fronted *wh*-element moves from the embedded clause into the main clause. In (2a), however, the moving

wh-element moves from one place in the main clause to another place, still in the main clause. The problem then seems to be with the fact that the *wh*-element attempts to move out of an embedded clause, across an intervening *wh*-element that was present in the same embedded clause. So, the main takeaway from (2) is that *wh*-movement out of embedded clauses leads to ungrammatical sentences when an intervening element is present. What I have described here is the concept of a syntactic island effect.

1.1 Syntactic Islands

Ross (1967) was the first to identify and describe syntactic islands. The imagery of naming these clauses islands, reflects the idea that someone cannot simply leave an island; they are stranded. Syntactic islands exist in many different shapes and forms, but they share a common theme: all syntactic islands represent clauses that somehow constrain elements from moving out of them. While these syntactic islands are all interesting in their own right, the current study will focus on *wh*-islands exclusively. *Wh*-islands are a variant of syntactic islands. Their specific constraint seems to be that they prevent *wh*-elements from moving out of an embedded clause while there is another *wh*-element present in that clause.

As a long standing tradition, syntactic islands have been considered, as the name suggests, to be a syntactic phenomenon. The idea there, is that a grammatical constraint is responsible for the depreciation of acceptability: The rule is violated, which leads to ungrammaticality. But, ever since the inception of the idea of syntactic islands, scholars have tried to find alternative theories for the unacceptability of island violations. This discussion has divided the field of linguistics for decades. Many linguists (e.g. Villata et al., 2016; Rizzi, 2017) still hold the position that a grammatical ban on island-extraction best accounts for the facts, while others (e.g. Hofmeister & Sag, 2010; Goodall, 2015) find evidence that other, extra-grammatical factors (such as limitations in working memory) could be responsible for the unacceptability of island extraction. Section 2 of this paper will cover this debate much more extensively.

So far, it is not exactly clear what exactly these factors or syntactic rules might look like, or how we should make a distinction between the proposals. One thing that is sure, is that the proposals each need to account for the facts found in the language. There is one strange behaviour of syntactic islands that plays a key role in the debate about which factors cause island effects. The behaviour I am referring to can be captured in the term *complexity*

effects. A strong proposal must be able to accommodate complexity effects. But what exactly is a complexity effect and why is it relevant to the discussion?

Complexity effects are a phenomenon of interest that manifests in island extractions. In the context of *wh*-islands, this comes down to the fact that island violations are graded to be more acceptable if the moving element (often also called a filler) is not a bare *wh*-word, such as *what* in (2), but instead a more complex lexical phrase, such as *which cake* in (3). Sentences with a complex-*wh* filler extraction like (3) still tend to receive low acceptability judgments, but are generally graded to be more well-formed than those in (2).

(3) a. ?Which cake did who eat _?
b. ??Which cake do you wonder [who ate _]?

This filler-complexity acceptability-ameliorating behaviour of islands is odd, because it requires islands to have a governing system that accounts for the whole spectrum of their acceptability judgments. Complexity effects have therefore spawned a whole host of new analyses on the nature of island constraints. Some of these proposals from both sides of the the seem to describe the available data on islands relatively well, but Beljon et al. (2021) presents a unique challenge. At first glance, some of their findings seem problematic for all currently available proposals, but they find an analysis using the syntactic principle of featural Relativized Minimality (fRM), which will be explained later. For now, we should note that the landing clause of the *wh*-filler will play a defining role.

1.2 The Present Study

The present paper will attempt to solve one piece of the puzzle that the data in Beljon et al. (2021) presented. Section 2 will explain the theoretical background for the debate on filler complexity and how an fRM proposal could potentially account for the facts captured in Beljon et al. (2021) The end of section 2 will postulate the hypotheses for the experiment presented in this study. Section 3 will explain in detail how the experiment was conducted, which materials were used, what the results of the experiment were and it will provide a detailed statistical analysis of that data. Section 4 will discuss the implications of the findings and how they align with the hypotheses. Section 5 will provide a conclusion to the study as a whole.

2. The Problem with Syntactic Islands

The general consensus is that island violation leads to bad sentences. A person is unlikely to utter sentences that violate island constraints, and island-violations are generally graded to be ungrammatical. However, certain manipulations (such as modulating bare and complex fillers, or adding context to the question) are shown to improve acceptability judgments of those sentences. There are two common types of accounts for the unacceptability of island violations. The first type of account supports the idea that formal syntactic rules in a grammar prevent people from interpreting sentences with island-violations as being viable constructions. These types of accounts are classified as grammar accounts.
The second type of account claims that sentences with multiple filler-gap dependencies tax the cognitive capabilities of the interpreter too much, which leads to the inability to process those sentences as interpretable. These arguments are classified as Reductionist or Processing/Working Memory accounts.

A major point of contention in the discourse on syntactic islands is the ameliorative effect of filler complexity. More complex fillers (4b) are usually graded to be more acceptable than bare fillers (4a).

(4) a. *What do you wonder [who ate ___]?

b. ?Which meal do you wonder [who ate ___]?

The amelioration of these types of sentences is relatively well researched, and this phenomenon has received numerous explanations. The difference in acceptability between (4a) and (4b) is the fact that all theories on syntactic islands should attempt to capture. The view that currently still holds the most sway is that syntactic islands behave according to a formal grammatical rule, because that is how they have been analysed historically. There are many theories for both the reductionist and the grammatical sides, that try to explain the unacceptability of island extractions, but this background section will focus on three of the proposals: Relativized minimality (a grammatical account), Processing (a reductionist account) and D-linking approaches.

2.1 Grammatical Accounts

The introduction shortly touched on the idea that grammatical constraints were responsible for island effects. Grammatical accounts rely on the idea that island effects represent filler-gap dependencies. When a constituent moves from one location to another, formal syntax has established that it leaves a gap at the location that it originated from. The element that moved

is then called the filler. The explanation for filler-gap dependencies that currently seems to be the fittest for capturing all the facts is a fRM, a variant of RM. How those work and how this theory has evolved will be discussed below.

Many grammatical accounts involve some version of Relativized Minimality, or RM for short (Rizzi, 1990). Relativized Minimality postulates that syntactic processes are inherently local: they are bound to apply within a limited structural domain. Often, locality principles force constituents to move to different positions in the sentence, so that all kinds of dependencies can be fulfilled locally. In Relativized Minimality, two constituents (X and Y) cannot be related to one another while a third constituent (Z) *intervenes* between them. That means, a constituent Z, that is comparable to the higher constituent X in the left periphery, interferes in the relationship between X and Y. See (5).

(5) . . . X . . . Z . . . Y . . .

RM proposes an explanation for the ungrammaticality of wh-island extractions. A *wh*-filler leaves an invisible trace at the gap site Y (5) when it is re-merges at position X (5). The filler and gap enter into a dependency: the interpreter of the sentence needs to understand that the filler came from the gap site to be able to reconstruct the sentence. But syntactic island extractions are ungrammatical because the relationship between the filler (5: X) and the trace (5: Y) is disrupted by the intervention of the other *wh*-element in the clause (Z). This happens because when X and Z are functionally comparable, the RM principle will interpret the filler-gap dependency as a relation between X and Z instead of a relation between X and Y. RM is a robust proposal, but in its current form, it does not account for the spectrum of different acceptability ratings that complexity effects introduced.

Friedmann et al. (2009) proposes a refined version of RM: featural Relativized Minimality (fRM). The important distinction between RM and fRM is that fRM can account for different degrees of acceptability (for example those brought on by complexity effects). It does so by explaining that the intervention effect of Z can be stronger or weaker, depending on how many morphosyntactic features it has in common with X. The fRM constraint is defined in (6).

(6) “Featural Relativized Minimality:

In . . . X . . . Z . . . Y . . .

A local relation is disrupted between X and Y when

a. Z structurally intervenes between X and Y
b. Z matches the specification in morphosyntactic features of X”

(from Villata et al, 2016; 78)

The degree of disruption is based on the ratio of features X and Z have in common. A higher the degree of correspondence of features between the X and Z leads to a stronger intervention effect. The intervention effect is strongest when the intervening element shares all features (7a: Identity) with the filler in the left periphery. The intervention effect (and thus the island effect) is weaker when they only share a subset of features (7b: Inclusion). See the contrast in (7), using the system of features as proposed in Friedmann et al. (2009).

(7) a. *What do you wonder [who ate __]? Identity
[Q] [Q]
b. ?Which meal do you wonder [who ate __]? Inclusion
[Q, N] [Q]”

According to the fRM framework, (4b) and is graded better than (4a), because *which meal* contains more features than *what*. This means that *who* has full featural overlap with *what*(Q : Q → Identity) and only partial feature overlap with *which meal*(Q,N : Q → Inclusion). The consequence of this fRM proposal is that complex wh-elements should be graded as generally better than bare wh-elements, because complex wh-elements have more features, so they are less likely to have full featural overlap (Identity) with the intervening wh-element The fRM proposal is fine-tuned to account for the difference in acceptability of bare- and complex *wh*-fillers. fRM will be a key component of the argument the present study attempts to present. Section 2.4 will elaborate further on this.

I believe that this section should introduce the difference between strong and weak islands as well. Strong islands do not allow any kind of constituent to be extraction from them. Weak islands disallow some extractions, but (partially) allow other extractions. If a language can be shown to have any type of amelioration in island extractions, then that language by definition has weak islands. The literature suggests that most languages have weak islands, because many languages allow some form of island violation amelioration. However, Beljon et al. (2021) found no convincing evidence that complexity amelioration reflected a grammatical constraint. This suggests that Dutch has strong islands. Languages

with strong islands cannot use complexity effects to account, unless they can explain in a different way why their islands are strong, as well as show a way to weaken those islands.

2.2 Reductionist Accounts

Reductionist accounts generally tend to argue that island violations are, in fact, perfectly grammatical sentences, but that their unacceptability arises from extra-grammatical factors; independently motivated effects that exists outside the grammar. In other words, grammar has nothing to do with why island-extractions are considered bad, but some other mechanism (such as a limit on processing capabilities) interferes with the comprehension of sentences that contain island violations.

The main argument of reductionist accounts I wish to outline is the following: Scholars have found that the reason that complexity effects ameliorate island extractions stem from factors related to working memory or processing capabilities, rather than any grammatical constraints. Some scholars suggest that lexical (complex) wh-phrases are processed slower, but the frontloaded effort spent on processing the filler then facilitates the retrieval process from short term memory at the gap, which is an overall decrease in the burden placed upon one's processing faculties. This decrease in processing cost subsequently leads to an increase in acceptability (Hofmeister & Sag, 2010; Hofmeister et al. 2013).

Hofmeister et al. (2013) looked at the effect of filler complexity on graded judgments in multiple wh-sentences in English. They found a significant complexity effect with the manipulation of processing variables. They reported the following filler complexity hierarchy:

(8) "bare-bare < which-bare ≤ bare-which < which-which" (taken from Rizzi, 2017, 254)

We can compare this to Villata et al. (2016), who also found a significant complexity effect in French, by manipulating filler variables of the syntactic structures. They found the following filler complexity hierarchy:

(9) "bare-bare = bare-which < which-bare < which-which" (taken from Rizzi, 2017, 254)

The notable difference between (8) and (9) is the inversion of acceptability between the *bare-which* and *which-bare* conditions. These results seem to contradict each other, but Rizzi (2017) found a way to reconcile both results under an adapted version of RM. This is adapted version of RM requires a very technical analysis that I will omit for the sake of brevity and the clarity of the argument. The argument relies on the fact that French and English have differences in overt and covert movement, but the main takeaway is that the RM

framework is capable of accounting for the different findings between the grammatical proposal of Villata et al. (2016) and the processing account of Hofmeister et al (2013).

2.3 D-Linking Arguments

There is another argument that, similarly to the complexity effect argument, could potentially play a pivotal role in providing evidence for either side of the island argument: D(iscourse)-linking. Making any claims about D-linking is outside the scope of the present study, since the experiment presented will not rely on D-linking evidence. However, since D-linking is a phenomenon that applies to syntactic islands and seems to interact on a similar domain as complexity effects, accompanied the fact that D-linking is brought up in many accounts of the island problem, it should prove valuable to have some understanding of the phenomenon of D-linking. This section will offer a quick overview.

Pesetsky (1987) proposed that complex *wh*-phrases like *which car* are linked to a discourse, in that they prompt an answer from among the elements that already exist in the discourse. The expectation in D-linking accounts is that both the speaker and the hearer both already know about a potential set of cars in (10a), but that this is not the case for (10b). The link to the discourse available in (10a) is claimed to weaken or sometimes even erase island effects.

(10) a. Which car did you buy?

b. What did you buy?

A key prediction by the D-linking argument is then that sentences with non-D-linked bare-*wh* island extractions like (10b) should improve to the same level of acceptability as (10a) when a context is added. Some scholars did indeed find evidence that inducing D-linking by adding context improves the acceptability of the bare-*wh* extractions (Goodall, 2015; Kush et al, 2019). However, some studies were unable to confirm D-linking effects (Villata et al., 2016). One study in particular stood out: Beljon et al. (2021). When they added context to bare-*wh* extraction, their acceptability depreciated. This is completely opposite of what D-linking theories expect, and it seems unlikely that a D-linking will be able to accommodate their findings.

Goodall (2015) outlined that there are three potential explanations for D-linking effects. The first proposal is that D-linking comes from semantic factors: it makes the interpretation of referents easier. The second proposal is that D-linking is a syntactic

phenomenon: D-linked fillers are reanalysed as fronted topics, which are immune to intervention effects such as the ones described in (f)RM accounts. A third and most interesting proposal is that D-linking helps with the working memory issues of processing filler-gap dependencies. Because linking a filler to the discourse requires more processing initially, the fillers survive better in the short-term working memory. This makes the fillers easier to access when processing the gap, which leads to an overall lower processing burden. The increased ease of processing is then theorised to lead to higher acceptability (Goodall, 2015).

The final proposal is interesting because it makes no distinction between syntactic island extractions and other filler-gap dependencies. This predicts that D-linking should be able to improve the acceptability of non-island environments. According to this working memory proposal, D-linking is a general phenomenon that is independent of the effects that govern the acceptability of syntactic islands. In this analysis of the D-linking phenomenon, D-linking will be unable to bear on the nature of syntactic islands.

Because of the inconsistencies with the induction of D-linking effects, as well as the inability for a D-linking approach to account for the data found by Beljon et al. (2021) and the uncertainty about where D-linking effects come from, the status of D-linking and its potential to inform about island effects is unclear at present.

2.4 Tackling the issues with Beljon et al.

Beljon et al. (2021) looked at main clause extractions of both bare-*wh* and complex-*wh* constituents from *wh*-islands. Initially they found no effect for filler complexity. In the second experiment they conducted, they added a context condition. This means that some of the experimental items were introduced after a short contextualisation of the question, and other questions were posed without added context. They discovered that adding context to island extractions affected the scores for complex-*wh* extractions positively, although the effect size was small. Strangely, the bare-*wh* extractions degraded in context. This means that the complexity effect arose not necessarily from the complex condition ameliorating, but also partly from the bare condition degrading. Therefore, they carefully concluded that there was no convincing evidence for complexity amelioration.

Dutch is classified as a V2 language. The V2 constraint is a rule that entails that Dutch main clauses must have a verb (V) as their second (2) constituent. This becomes relevant when you look at the features that trigger syntactic movement. Both Dutch and English share

the behaviour that *wh*-elements are triggered to move to clause-initial position. In other words, *wh*-elements are fronted. This *wh*-fronting effect is consistent across both main- and embedded clauses for both languages. So far, it seems that Dutch and English treat *wh*-questions the same. There is, however, one key difference: The V2 constraint of Dutch.

Featural Relativized Minimality (fRM) predicts that Inclusion should be graded better than Identity, and that Identity entails strong islands.

Beljon et al, (2021, 17) postulate that some unknown feature [F] (perhaps for front?) is added to the C-position of Dutch main clauses, due to the V2 constraint requiring the first constituent in the clause to be a certain type of element. This added Feature [F], makes it so that the C-position probes for an element with the feature [Q, F], which can be fulfilled by both complex and bare fillers in Dutch, since they are both elements that can take the C-position, and are both [Q]uestion elements. This means that both bare and complex extractions lead to a maximal intervention effect with the intervener (Identity) See (11).

(11) a. * Wat$_i$ vraag jij je af wie _$_i$ gebakken heeft? Identity
[Q, F] [Q, F]
'What do you wonder who baked?'

b. *Welke traktatie$_i$ vraag jij je af wie _$_i$ gebakken heeft? Identity
[Q, F] [Q, F]
'Which treat do you wonder who baked?'

There is one final key factor on which the fRM proposal of Beljon et al. (2021) is contingent. Goodall (2015) found that the complexity effects in English also apply to non-island conditions. Beljon et al. (2021) also found this behaviour for Dutch in their context conditions. If island complexity effects can indeed improve non-island sentences, then it follows that complexity effects are not a grammatical phenomenon. The problem with that is that, if complexity effects are indeed extra-grammatical, their ameliorative potential cannot bear on the grammatical nature of the island constraints themselves. Future research should investigate whether complexity effects can improve non-islands. If this can indeed be shown definitively, a reanalysis is necessary.

2.5 The Present Study

2.5.1 Standards of Evidence

It is difficult to ascertain which proposals hold up under scrutiny, since many accounts have been based on informally collected data. Consequently, attempts have been made to criticise flawed assumptions, methods, techniques and models (cf. Philips 2011; Sprouse et al., 2013; Sprouse et al., 2016) This current study uses testing methods similar to Beljon et al. (2021) and attempts to adhere to the guidelines and standards of evidence that Sprouse et al. (2016) recommended for formal collection of graded acceptability judgment data. This will be elaborated upon further in the experimental design in section 3.

2.5.2 Goal of the present study

The analysis in Beljon et al (2021) presents two sides of a puzzle to investigate. The present paper focuses on English, a language without V2, to hopefully show that English has a complexity effect that stays consistent through both main clause island extractions and embedded clause island extractions. A sister project of this present article is currently investigating whether it is indeed the case that Dutch embedded island extractions can show complexity amelioration while main clause island extractions do not. The comparison of these two papers can rule out whether the V2 constraint is responsible for the existence of strong islands in Dutch, or whether a new analysis is necessary to capture the facts. Future research shall show which of the two possibilities is borne out.

For the comparison between the two papers, It is necessary that the present study finds a consistent effect for filler complexity in both main clause island extractions and embedded clause island extractions, so that the putative result of the Dutch embedded clauses research can isolate V2 as the only formal difference between the situation in Dutch and the situation in English. To this end, this experiment adopts an EMBEDDEDNESS variable.

(12) a. **What** do you wonder [who ate ___]?
b. Does John know [**what** you wonder [who ate _]]?

The EMBEDDEDNESS variable reflects the syntactic structure that the experimental items in section 3 will have. In the *Main Clause* condition, the *wh*-element will be extracted from the syntactic island in the embedded clause, after which it resides at the front of the main clause. In the *Embedded Clause* condition, an additional embedded clause is introduced as the

landing site for the wh-element, so that the extracted *wh*-element exists in an embedded clause rather than in the main clause. See (12) for an illustration of this rather unwieldy explanation.

2.6 Hypotheses

On its own, this experiment has the goal of showing a consistent complexity effect for both island extractions that extract to main clauses and island extractions that extract to an embedded clause. This alone would be valuable data, as it would substantiate the position that English has weak islands, and that the islands remain weak regardless of where the extracted filler moves to. This will either show a difference or similarity between English main clauses and embedded clauses as predicted by the analysis in Beljon et al. (2021) and in the greater discourse, the study will provide contrast for research into other languages to verify predictions that were made by this study and Beljon et al. (2021)

Thus, the following hypotheses are postulated:

A. Complexity of the moving *wh*-element has a significant ameliorative effect on *wh*-island violations.
B. The complexity effect is consistent in both embedded and main clauses.

3 Experiment

3.1 Participants

Sixty-two native speakers of English (30 female, 30 male, 2 non-binary, mean age = 23.9 years, range = 18-44 years) participated in the survey entirely voluntarily and without recompense. For more detailed data about the participants, see Appendix 2. Participants were recruited through personal communication, social media platforms and various online community messaging boards. All participants had expressed informed consent for their participation in this study. Participants were anonymised and their data was stored in accordance with the Radboud University data storage guidelines.

3.1.1 Excluded individuals

Fifteen individuals did not finish the entire questionnaire. Consequently, their answers were deleted and their data was omitted from this paper. Individuals who attempted to participate in the survey, but did not give consent, as well as individuals who reported to not be native speakers of English, and finally, individuals who reported having language or language-development impairments (e.g. dyslexia) were respectfully excluded from the experiment. This led to two additional individuals being excluded from participation.

3.2 Design and Materials

3.2.1 Design Goals

In the interest of comparability between studies, the present study has strived to stay formally comparable to the design of Beljon et al. (2021) as well as possible. Additionally, it has strived to adhere to the guidelines on formal linguistic research posited in Sprouse et al. (2016) in order to uphold the standards for formal linguistic research. This means that the present experiment will also use 7-point Likert scale graded acceptability judgments with a high number of experimental items and conditions. Additionally, measures have been taken to reduce potential sources of bias in the data, each instance of which will be elaborated upon further, whenever they first appear throughout the rest of section 3. A final important design goal for this paper was to provide a clear formal description of the experiment, for the sake of reproducibility and transparency.

3.2.2 Design of the Experiment

The experiment has a 2x2 design (within-subjects and within-items), crossing the factors EMBEDDEDNESS (island extraction to main clause vs. island extraction to embedded clause) and FILLER COMPLEXITY (Bare *wh* vs. Complex *wh*). The materials consisted of 16 experimental items (with 4 conditions for each item) and 40 filler items (see Appendix 1), for a total of 56 items.

The experimental items were divided over four counterbalanced lists of questions, following a Latin square design, in order to minimize any bias that might arise from the order in which the conditions are presented. Each of the 16 experimental items was only displayed to each participant in one of its conditions. All 40 filler items were added to every list. This means that each participant saw each of the four conditions exactly four times, and all 40 filler items once, for a total of 56 items per list of questions. All items were pseudo-randomly divided into 8 blocks of 7 questions. Within and between blocks, experimental items were never followed or preceded by other experimental items or subject island filler items in order to prevent syntactic priming. The blocks were also presented to participants in a pseudo-randomised order, but the first three questions presented to participants were always filler items: one from each of the different filler qualities. This was done to acquaint the participants with the assignment and to calibrate their expectations with regards to the range of grammaticality they could expect to see in the experiment.

3.2.3 Experimental Items

Each experimental item was an embedded question sentence, which was displayed in any one of the four different experimental conditions (see table 1). The element that was extracted from a syntactic island was inanimate in every condition, with *what* as the bare *wh*-filler and *which [inanimate noun]* as the complex *wh*-filler in each experimental item. These choices were made to minimize ambiguity between subject and object, which may arise when two or more *wh*-elements are present in a sentence (Donkers et al., 2013) as well as to stay formally comparable to Beljon et al. (2021).

In all conditions, the extracted filler was moved from the syntactic island to the front of the clause that embeds the island (so it moves 1 layer up). For the -Embedded conditions, this means the filler moves to the front of the main clause, whereas in the +Embedded conditions, the filler moves to the front of the first embedded clause that is structurally directly above the island. The subject of the clause where the filler ended up was *you* in every

condition and the question was always introduced by the matrix verb *to wonder*. In the +Embedded structures (table 1, C and D), the main clause that introduced the embedded *wh*-island question always used the verb *to know*, with different animate noun phrases as the subject for every item, but mostly distinct proper names. Table 1 provides samples of these experimental items.

	Experimental condition	Sample sentence
A.	**-Complex** -Embedded	**What** do you wonder who ate _?
B.	+**Complex** -Embedded	**Which fruit** do you wonder who ate _?
C.	**-Complex** +Embedded	Does John know **what** you wonder who ate _?
D.	+**Complex** +Embedded	Does John know **which fruit** you wonder who ate _?

Table 1: Experimental conditions

3.2.4 Filler Items

A total of 40 filler items were created. They covered the full spectrum of acceptability for question sentences. Out of the 40 filler items, 16 are classified as acceptable (good filler items), 12 as intermediately acceptable (medium filler items) and 12 as strictly unacceptable (bad filler items). Embedded questions in filler items were always introduced with different matrix verbs than those used in the experimental items, in order to avoid any bias that might arise from the effects of priming. Following the expectation that the experimental items would be graded mostly unacceptable or intermediately acceptable, this distribution of filler items aims to somewhat smooth out the number of items across the acceptability spectrum, as advocated for by Sprouse et. al (2016).

Most of the filler item categories were adopted from Beljon et al. (2021), but verb clusters and preposition mistakes are not generally graded to be intermediately grammatical in English. Thus, the present study opted for the insertion of two new categories of medium-

acceptability fillers, namely semantically strange imperfect predicates, and semantically constrained passive constructions, which are found to be intermediately grammatical in English (cf. Ambridge et al., 2015; Bach, 1986).

The good filler items (16 total) consisted of three different types of conventionally acceptable sentences. Table 2 illustrates the three types of good filler items with sample sentences.

Filler Type (No. items)	**Sample Sentence**
Yes/no questions without islands (6 questions)	*Do you think that he ordered the wine?*
Wh-questions with an adjunct phrase (6 questions)	*When do you suppose she picked up her driver's license?*
Direct object extraction without islands (4 questions)	*What do you think Paul ate?*

Table 2: Good filler items

The medium filler items (12 total) consisted of four different types of intermediately acceptable sentences. Each of the four filler types are depicted in table 3 and illustrated with sample sentences.

Filler Type (No. items)	**Sample Sentence**
Semantically strange imperfective predicate (3 questions)	*Do you think Jack went on an adventure often this morning?*
Semantically constrained passive (3 questions)	*Who do you think this road was walked by*
Partial wh-movement (3 questions)	*What do you expect who gave the present?*

Filler Type (No. items)	Sample Sentence
Wh-copying (3 questions)	*Who do you suspect who stole the watch?*

Table 3: Medium filler items

The bad filler items (12 total) consisted of three different types of generally ungrammatical sentences. Table 4 shows the different types of bad filler items and clarifies them with examples.

Filler Type (No. items)	Sample Sentence
Subject Islands (4 questions)	*Who was the story of boring?*
Agreement mistakes (4 questions)	*Do you suspect that he have caused the destruction?*
Word salad (4 questions)	*Doubt the pike that you caught he have?*

Table 4: Bad filler items

These filler items served three purposes. Their first purpose was to counteract a potential response equalization bias: Because the filler items had many different aspects for why they could be considered good or bad, participants should be less likely to equalize their responses for very similar sentence structures. The second role of the fillers was to check whether or not participants were filling in the questionnaire seriously. This means that the bad filler items doubled as control items. If they were rated highly by participants too often, the results of those participants were treated as unreliable and omitted from the analysis (more on this in section 3.4 Data Analysis). Finally, due to limitations and time pressure, this research could not investigate 3-way interactions. Islandhood was not included as a formal variable and the present research was carried out under the well-substantiated assumption that an island-effect exists. Reaffirming the existence of such an island-effect with a separate experiment was unfortunately outside the scope of this paper. Nevertheless, to somewhat substantiate the existence of island effect, the good filler items with direct object extractions will be used to informally determine whether such an effect is indeed also present in the

sample population of this experiment. The direct object extractions are expected to be scored higher than the experimental items, which contain island extractions, but these results will not be subjected to any formal statistical analyses (more on this in section 3.5.2 Islandhood).

3.3 Procedure

In the experiment, participants were asked to perform off-line graded acceptability judgments. Before starting the experiment, participants were given the instructions to imagine that the items were spoken to them by native speakers they know well, such as a good friend or family member. The instructions explained that their task was to grade sentences. They were told to rely on their first impression and to grade based on how natural the sentence sounds to them. A caveat was provided in the explanation, wherein participants were dissuaded from trying to answer the questions posed in the items. They were also told that they should not factor in the degree to which the question is informative in their acceptability judgments.

The experiment was conducted using Qualtrics XM (Qualtrics, Provo, UT) survey software (Radboud University License). Items were presented in the centre of the screen and one at a time. Below each item, a 7-point scale was displayed. The text “Very poor” was displayed on the lefthand side of the scale, while “Very good” was displayed on the righthand side of it (see figure 1 below for a sample item). Participants were required to submit an answer for each item and they then had to click the arrow button to proceed to the next item. Participants could not go back to view or change their ratings of previous items.

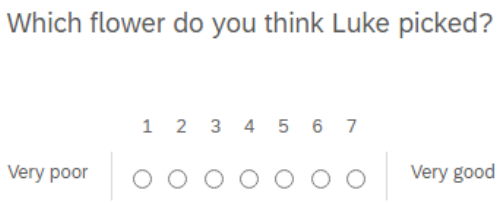

Figure 1: Sample Item (Good filler 16: direct object-extraction)

The final question in the survey thanked the participant for their participation and asked them to explain briefly what they thought the experiment tried to research. If participants were aware of the goal of the research, they had the opportunity to manipulate their answers to skew the results. Therefore, any participant that answered correctly should be excluded from the analysis. The next section will explain whether it was necessary to exclude any participants, and go over the data that was derived from the experiment.

3.4 Data Analysis

Before the data was analysed, certain exclusion criteria were checked. The bad filler items (see table 4) were used as a control variable. The following exclusion criteria were set on the bad filler items: no more than 4 out of the 12 bad filler items were allowed to be rated 4 or higher on the Likert scale. Additionally, at most 2 out of the 4 word salad filler items should be rated 4 or higher. Three participants had to be removed from further analysis based on these criteria. The answers given to the final question about the aim of the research were reviewed, and no participant mentioned (in more or less specific terms) the acceptability of syntactic island extraction or the amelioration thereof. One participant mentioned embedded clauses, but in the context of unnatural speech generation, which was judged to be far enough removed from the aim of this present research and was therefore admitted for analysis. In conclusion, no further participants were excluded based on that criterion. Fifty-nine participants remained for the final statistical analysis.

Raw acceptability scores were converted mean scores for each participant and for each condition. This means that each participant had a mean score for each condition. A two-way repeated measures ANOVA was performed on these data to evaluate the effects of EMBEDDEDNESS and FILLER COMPLEXITY on the mean scores. The assumptions for this analysis were checked beforehand. One extreme outlier was found. The outlier was allowed to remain in the data because the nature of the data allows for large differences between acceptability scores, but for the sake of the analysis the data was adjusted by taking the log() of the means for assumption checks. After the adjustments, no extreme outliers were found. Normality was checked with the Shapiro-Wilk test, which failed to show that the data was likely to be normally distributed($p<0.05$). The log()-adjusted data was also checked for normality with the same test, which also failed to show normality($p<0.05$). The adjusted data has a skewness of 0.28, which is in the acceptable range for normality (-1.0 to 1.0), but the kurtosis was sizable enough to the point where it was unable to predict a normal distribution of the data ($2.59>1.0$). Finally a Q-Q plot was made of the log()-adjusted data and visually inspected (see Figure 2). While there are some datapoints outside of the grey confidence range, the outliers appear on both sides of the reference line and most of the adjusted means of measurements are approximately on the line, which suggests that the data leans towards a normal distribution. It is, however, not unequivocal evidence of normality. The judgment was made that due to the high volume of measurements, the data is most likely robust enough to hold up, despite the violation of the normality assumption. This judgment carries the

implication that conclusions have to be drawn carefully, and that those conclusions would be better substantiated by collecting even more data. Sphericity was checked internally, with Maunchly's test of sphericity, by the R function anova_test() (from the rstatix package). If any factors violated the sphericity assumption, the R function get_anova_table() (also from rstatix package) automatically corrected for them with the Greenhouse-Geisser sphericity correction.

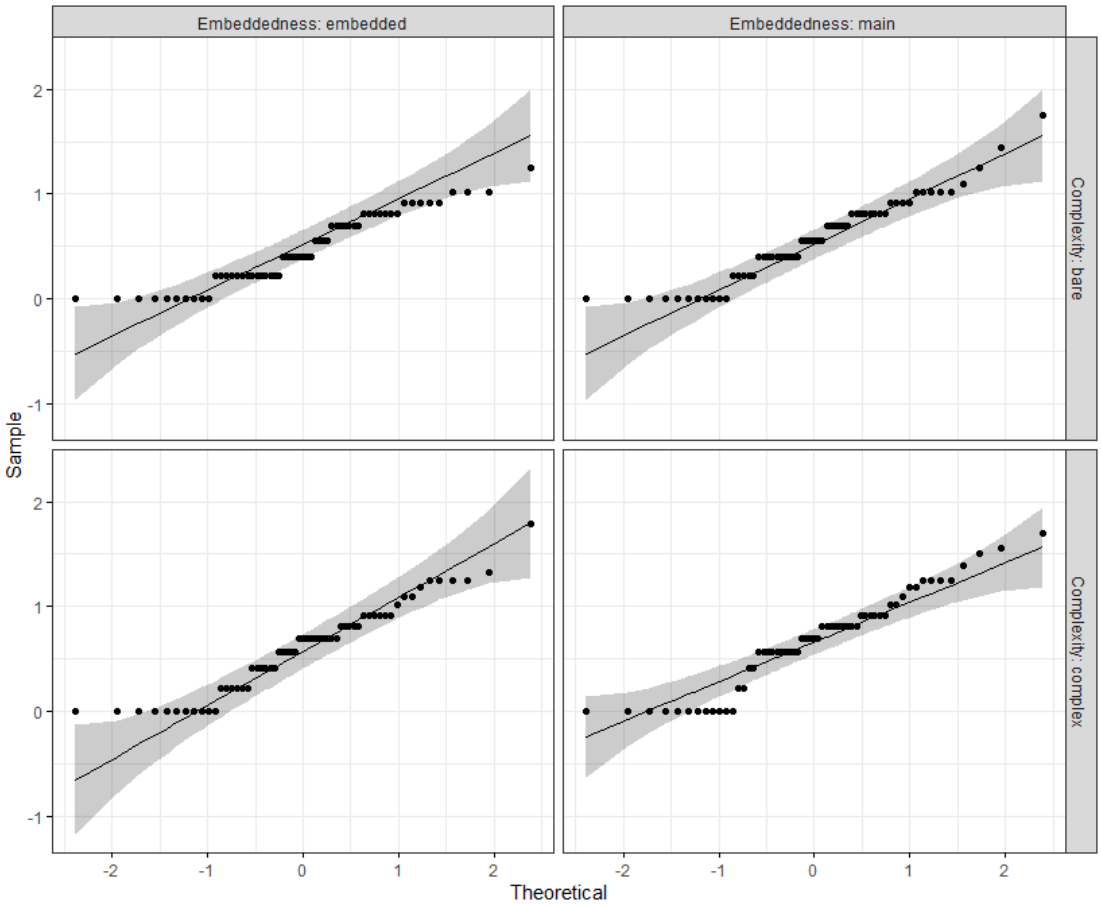

Figure 2: Q-Q plots of adjusted means

3.5 Results

Table 5 shows the mean acceptability scores per experimental condition with their corresponding standard deviations. No significant two-way interaction effect was found between EMBEDDEDNESS and FILLER COMPLEXITY on acceptability scores ($F(1,58) = 0.14$, $p = 0.71$). Two significant main effects were found with the ANOVA. EMBEDDEDNESS was significant for predicting scores in the way that sentences with the *main clause* factor were rated significantly better than sentences with the *embedded clause* factor ($F(1,58) = 8.17$, $p =$

0,0060) and FILLER COMPLEXITY was significant in the way that sentences with the *bare-wh* factor were rated significantly lower than sentences with a *complex-wh* factor.

	Bare-*wh*		**Complex-*wh***	
	M	SD	M	SD
Main clause	1.92	0.86	2.17	1.01
Embedded clause	1.70	0.60	2.00	0.93

Table 5: Mean acceptability scores and standard deviations per condition

Partial eta-squared tests were applied to both significant main effects to determine their effect size. The effect size of the embedded condition was medium (pes = 0,12), while for the complexity condition, much more of the variance could be attributed to that factor. The effect size for complexity was almost twice the threshold value for an effect to be considered large (pes = 0.27). The two significant main effects were followed up with pairwise comparisons with Bonferroni p-value corrections. Both pairwise comparisons were found to be statistically significant: Complexity(p = 0,000014) and Embeddedness (p = 0.004).

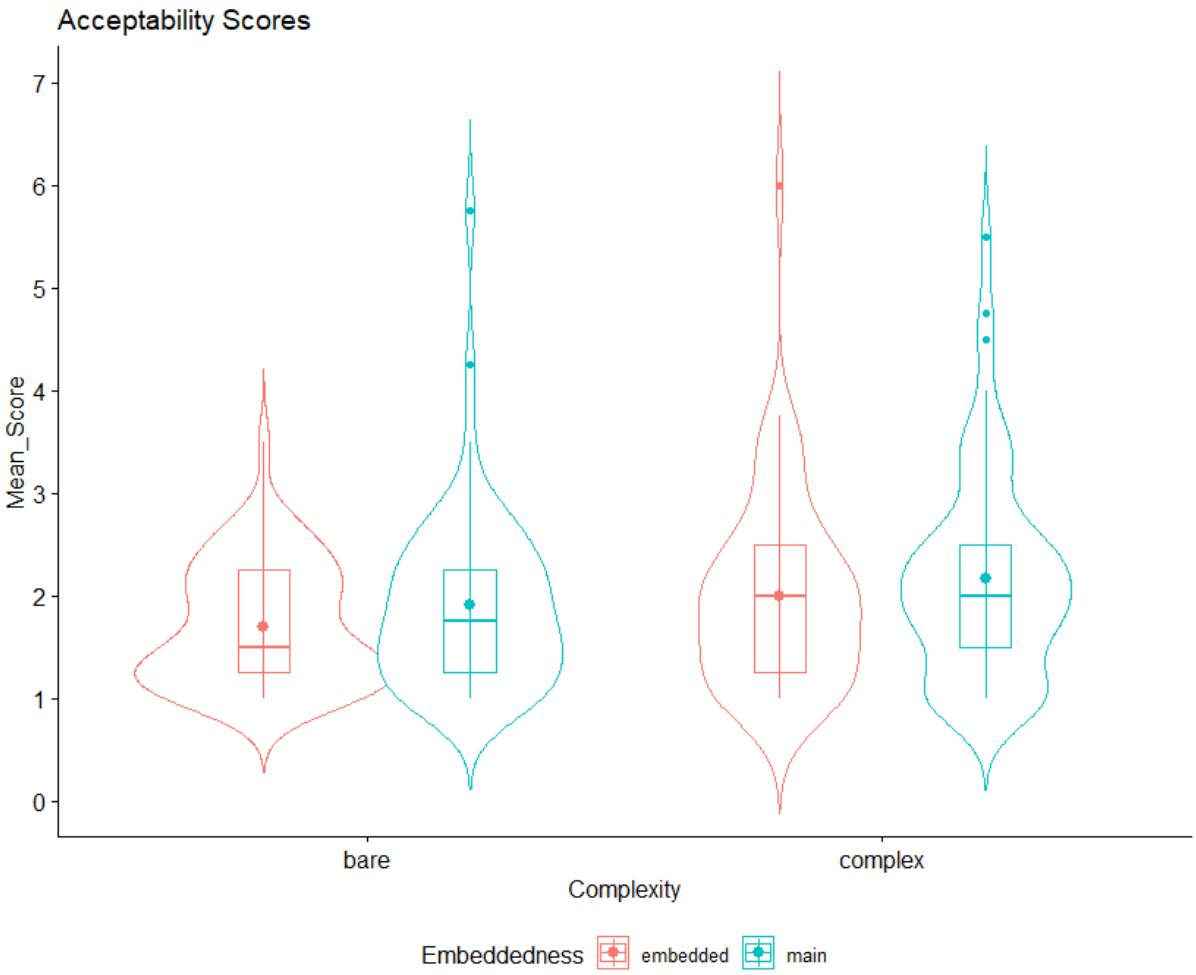

Figure 3: Acceptability score boxplots/violins with mean SD dots

In figure 3, the data of the study has been visualised. The width of the violin shapes represent frequency measurements of the mean scores. The boxplot with the vertical line show the 95% confidence interval, which reaches from the upper bound to the lower bound of the vertical line. The box of the boxplot represents everything within one standard deviation from the mean, which is represented by the horizontal line in the boxplot. The dot displays the mean standard deviation.

3.5.2 Islandhood

As explained at the end of section 3.2.4 (the section on filler items), limitations and time constraints unfortunately prevented this small scale experiment from including ISLANDHOOD (extraction from an island vs. extraction from a non-island) as a formal variable for a 3-way analysis. Because of the limited scope, the experiment had to be carried out under the assumption that island-effects do indeed exist in English. Fortunately, as the literature review section of this paper explained, island-effects are well researched, and almost ubiquitously found to exist. However, to further substantiate our already well-supported assumption, an

informal comparison could be made between some of the filler items and some of the experimental items. The filler items contained sentences with direct object-extractions (e.g. section 3.2.4, table 4; section 3.3, figure 1), which are structurally very comparable to the main clause *wh*-island extraction sentences in the experimental items (e.g. section 3.2.3, table 1, A and B). The assumption that an island-effect exists predicts that direct object-extractions are unproblematic and acceptable, but island extractions should be graded a lot worse.

Just like the literature predicted, the direct object-extraction filler items were graded notably better than their *wh*-island extraction counterparts, the A and B conditions of table 1 (i.e. mean score of 6.60 for non-islands vs. mean score of 2.04 for main clause *wh*-island violations). Although no formal statistical analysis has been done to formally confirm or reject the existence of island-effects, this informal comparison suggests that regular direct object-extractions to main clauses are graded to be well-formed, whereas *wh*-island extractions to main clauses are graded to be poor. This implies that there is indeed an island-effect, but because of the informal and non-statistical nature of the comparison, this inference should be taken as no more than a suggestion to aid the well-supported assumption that island-effects exist in English.

4 Discussion

4.1 The Data

This study researched the effect of EMBEDDEDNESS and FILLER COMPLEXITY on the acceptability of *wh*-island violations in question sentences in English. A significant effect was found for both conditions, although filler complexity had a larger influence on the acceptability scores than embeddedness. Complex gap fillers had a significant ameliorative effect on island violations, while the embedding of the island extraction questions had a significant degrading effect. Overall, the effect of complexity seems consistent between main clauses and embedded clauses, which is predicted by the hypotheses of this experiment. The statistical value of the results of this experiment should be interpreted with some caution. A potential criticism of this study could be based on the type of statistical analysis that was used to obtain and analyse the experimental data, because Likert scale acceptability judgment data is not usually assumed to be normally distributed (cf. the discussion section of Goodall, 2015 for examples of possible concerns). This makes sense, because subjective judgments are difficult to encode as formal quantitative data. This experiment has attempted do so to the best of its ability. In doing so, this paper has aimed to take part in the collective effort to raise the

bar for empirical research in linguistics spurred on by Sprouse et al. (2016), among others. Every effort was taken to reduce bias and increase the quality of data, but there was one oversight. The experiment could perhaps have been aided by transforming the raw acceptability scores to z-scores per participant. This transformation to z-scores could eliminate the risk of scale biases between participants. Examples of such biases include participants using a particularly large or small range of ratings (one participant might only answer with 3-5, while another might give ratings on the full 1-7 scale), or sticking notably to one side of the scale (e.g. giving mostly high ratings). These biases are eliminated by the z-score conversion because z-scores reflect the standard deviation of the raw ratings of a participant to their total mean rating. This means that their ratings can then be compared on a standardized scale. Unfortunately, it was not feasible to do this analysis post hoc, due to time constraints. The data is still likely robust enough to have statistical value, but future experiments should take this on board. Therefore, I believe that, after careful consideration, both hypotheses should be accepted, albeit with caution. So now, let us continue with an interpretation of the results.

The FILLER COMPLEXITY variable had a very large effect size, and therefore affected scores much more than the EMBEDDEDNESS variable did. The finding of a complexity effect of the island violating *wh*-element (bare-bare < which-bare) are in line with the expectations other research that investigated potential complexity effects in *wh*-islands and other cases of multiple *wh*-movement (e.g. Hofmeister & Sag, 2010; Goodall, 2015; Villata, et al., 2016) as well as theoretical accounts that explain filler complexity effects, such as featural Relativized Minimality (Rizzi, 2017, Beljon et al. 2021).

It is interesting that Goodall (2015) also found complexity-related amelioration effects outside of island contexts for English, since this is not predicted by any current grammatical theories. Further research should look at the 3-way interactions between islandhood, embeddedness and complexity.

The significant effect that was found for the EMBEDDEDNESS condition shows that the embedded sentences with *wh*-island extraction are generally rated to be less acceptable than main clauses with a *wh*-island violation. The fact that no interaction effect was found means that the influence of EMBEDDEDNESS on the score a sentence received was consistent between both bare and complex *wh*-islands. At first glance, it appears to be evidence for processing accounts that more complex sentences with multiple levels are rated worse because they overload an already stressed system even further. However, no analysis was carried out for

the effect of double embedded clauses outside of island-extraction clauses. According to processing accounts, it is to be expected that more complex sentences lead to slightly lower scores and the data reported in this study reflects that across the board. Future research could try to verify this empirically by contrasting the ratings of multiple levels of embedded sentences.

While *wh*-islands are among the best studied, there are of course other types of syntactic islands and there is always the challenge of making generalisations for all island types after investigating only one type. It is not unthinkable that islands behave similarly across the different types and that there is one single grammatical model that accounts for all of them (cf. Sabel, 2002). Conversely, it could be the case that different types of island are subject to different sets of constraints(cf. Sprouse et al., 2016). More research is necessary to draw conclusions that can generalize the findings to the whole set of syntactic islands.

The present study is a small scale experiment and as such lacks the scope, time and resources to find answers to many of the questions that presently go unanswered about variation in island acceptability. Ideally, one additional variable or experiment, such as crossing the factor islandhood(extraction from an island vs extraction from a non-island) or context would have been added to allow for a more informative comparison between factors that affect acceptability and would decrease the number of assumptions that the experiment had to make at present. Nevertheless, this experiment still found useful data for comparison with similar research on different languages and variables, and makes clear predictions about future research: English behaves according to the predictions made in Beljon et al. (2021), and if future research can show Dutch to behave as predicted as well, there is strong evidence for their hypothesis that the V2 constraint in Dutch is responsible for the lack of a complexity effect. If so, there would be no reason to assume other, similar languages would behave differently from one-another. Future research will show whether or not Dutch behaves as predicted, too, or if a new analysis is needed to capture the facts.

5 Conclusion

The ameliorative effect of complexity was consistent between main clause island extractions and embedded clause island extractions. The hypotheses of the experiment have therefore been successfully confirmed. This result provided valuable insights, both because it showed that English, as a language without a V2 constraint does indeed feature complexity amelioration, and because it provided a piece of evidence in favour of a fRM analysis of Beljon et. al (2021).

Acknowledgments

I would like to thank Olaf Koeneman for his guidance in this project, as well as Veerle Wilms for assisting me with the Qualtrics survey software and Susanne Brouwer for consultation about statistical analyses, and finally Jill Ostermeier for helpful discussions and insight about this study. All errors remain my own.

Bibliography

Ambridge, Ben, Bidgood, Amy, Pine, Julian M., Rowland, Caroline F. & Daniel Freudenthal (2015) "Is Passive Syntax Semantically Constrained? Evidence From Adult Grammaticality Judgment and Comprehension Studies". *Cognitive Science* 40. 1435-1459. doi: 10.1111/cogs.12277

Bach, Emmon. 1986. "The algebra of events." *Linguistics and Philosophy* 9(1). 5–16. doi: 10.1002/9780470758335.ch13

Beljon, Maud, Joosen, Dennis, Koeneman, Olaf, Ploum, Bram, Sommer, Noëlle, Swart, de, Peter & Veerle Wilms (2021) "The effect of filler complexity and context on the acceptability of *wh*-island violations in Dutch." *Linguistics in the Netherlands* 38 1:4-20 doi: 10.1075/avt.00047.bel

Donkers, Jantien, Hoeks, John & Laurie Stowe (2013) "D-Linking or set-restriction? Processing which-questions in Dutch." *Language and Cognitive Processes*28 (1–2): 9–28. doi: 10.1080/01690965.2011.566343

Friedmann, Naama, Belletti, Adriana & Luigi Rizzi. 2009. "Relativized relatives: types of intervention in the acquisition of A-bar dependencies." Lingua 119 (1): 67–88. https://doi.org/10.1016/j.lingua.2008.09.002

Goodall, Grant (2015) "The D-linking effect on extraction from islands and non-islands." *Frontiers in psychology*5: 1493. doi: 10.3389/fpsyg.2014.01493

Hofmeister, Philip & Ivan A. Sag (2010) "Cognitive constraints and island effects." *Language* 86 (2): 366–415. doi: 10.1353/lan.0.0223

Hofmeister, Philip, Casasanto, Laura S. & Ivan A. Sag (2013). Islands in the grammar? Standards of evidence. In J. Sprouse & N. Hornstein (Eds.), *Experimental Syntax and Island Effects* (pp. 42-63). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139035309.004

Kush, Dave, Lohndal, Terje & Jon Sprouse. 2019. "On the island sensitivity of topicalization in Norwegian: An experimental investigation.: Language 95 (3): 393–420. https://doi.org/10.1353/lan.2019.0051

Pesetsky, David (1987) "Wh-in-Situ: movement and unselective binding". *The Representation of (in)Definiteness* ed. By Eric Reuland and Alice ter Meulen, 98–129. Cambridge, MA: MIT Press.

Phillips, Colin (2013). "On the nature of island constraints I: Language processing and reductionist accounts." In J. Sprouse & N. Hornstein (Eds.), *Experimental Syntax and Island Effects* (pp. 64-108). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139035309.005

Rizzi, Luigi. (1990). "Relativized minimality." Cambridge: Cambridge University Press.

Rizzi, Luigi (2017) "Comparing extractions from wh-islands and superiority effects." *Wiener Linguistische Gazette* 82: 253–261. [Google Scholar]

Ross, J.R.(1967). "Constraints on Variables in Syntax" *Doctoral dissertation.* Massachusetts Institute of Technology. Norwood, NJ: Ablex. Published as *Infinite syntax* (1986).

Sabel, Joachim (2002) "A minimalist analysis of syntactic islands." *The Linguistic Review* 19: 271–315. doi: 10.1515/tlir.2002.002

Sprouse, Jon, Wagers, Matthew W. and Colin Philips (2013) "Deriving competing predictions from grammatical approaches and reductionist approaches to island effects." *Experimental Syntax and Island Effects*: 21-41 https://doi.org/10.1017/CBO9781139035309.003

Sprouse, Jon, Caponigro, Ivano, Greco, Ciro & Carlo Cecchetto (2016) "Experimental syntax and the variation of island effects in English and Italian." *Natural Language and Linguistic Theory* 34: 307–344. 10.1007/s11049-015-9286-8

Sprouse, Jon, Carson T. Schütze, & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001-2010. *Lingua* 134: 219-248. doi: https://doi.org/10.1016/j.lingua.2013.07.002

Villata, Sandra, Rizzi, Luigi, & Julie Franck (2016) "Intervention effects and relativized minimality: new experimental evidence from graded judgements." *Lingua*179: 76–96. doi: 10.1016/j.lingua.2016.03.004

Appendix 1

Experimental Items (16 items)

Conditions

A = Main clause + bare wh-phrase	**-embedded, -complex**
B = Main clause + complex wh-phrase	**-embedded, +complex**
C = Embedded clause + bare wh-phrase	**+embedded, -complex**
D = Embedded clause + complex wh-phrase	**+embedded, +complex**

1.

A. What do you wonder who ate?
B. Which fruit do you wonder who ate?
C. Does John know what you wonder who ate?
D. Does John know which fruit you wonder who ate?

2.

A. What do you wonder who watched?
B. Which show do you wonder who watched?
C. Does Mary know what you wonder who watched?
D. Does Mary know which show you wonder who watched?

3.

A. What do you wonder who chose?
B. Which course do you wonder who chose?
C. Does Anna know what you wonder who chose?
D. Does Anna know which course you wonder who chose?

4.

A. What do you wonder who bought?
B. Which plant do you wonder who bought?
C. Does Pete know what you wonder who bought?
D. Does Pete know which plant you wonder who bought?

5.

A. What do you wonder who baked?
B. Which treat do you wonder who baked?
C. Does Michael know what you wonder who baked?
D. Does Michael know which treat you wonder who baked?

6.

A. What do you wonder who used?
B. Which paint do you wonder who used?

C. Does Gary know what you wonder who used?
D. Does Gary know which paint you wonder who used?

7.
A. What do you wonder who drove?
B. Which car do you wonder who drove?
C. Does Susan know what you wonder who drove?
D. Does Susan know which car you wonder who drove?

8.
A. What do you wonder who crafted?
B. Which art project do you wonder who crafted?
C. Does Tim know what you wonder who crafted?
D. Does Tim know which art project you wonder who crafted?

9.
A. What do you wonder who folded?
B. Which paper hat do you wonder who folded?
C. Does he know what you wonder who folded?
D. Does he know which paper hat you wonder who folded?

10.
A. What do you wonder who practiced?
B. Which sport do you wonder who practiced?
C. Does Amber know what you wonder who practiced?
D. Does Amber know which sport you wonder who practiced?

11.
A. What do you wonder who wrote?
B. Which article do you wonder who wrote?
C. Does Vera know what you wonder who wrote?
D. Does Vera know which article you wonder who wrote?

12.
A. What do you wonder who visited?
B. Which theme park do you wonder who visited?
C. Does Thomas know what you wonder who visited?
D. Does Thomas know which theme park you wonder who visited?

13.
A. What do you wonder who sang?
B. Which song do you wonder who sang?
C. Does she know what you wonder who sang?
D. Does she know which song you wonder who sang?

14.

A. What do you wonder who replaced?
B. Which lightbulb do you wonder who replaced
C. Does Lisa know what you wonder who replaced?
D. Does Lisa know which lightbulb you wonder who replaced?

15.

A. What do you wonder who made?
B. Which dish do you wonder who made?
C. Does Rick know what you wonder who made?
D. Does Rick know which dish you wonder who made?

16.

A. What do you wonder who listened to?
B. Which song do you wonder who listened to?
C. Does your friend know what you wonder who listened to?
D. Does your friend know which song you wonder who listened to?

Good Fillers (16 items)

Yes/no questions without islands

1. Do you think that she bought the bookcase?
2. Do you expect that she played Monopoly?
3. Do you doubt that she hired the van?
4. Do you think that he ordered the wine?
5. Do you doubt that he trimmed the bonsai?
6. Do you expect that he found the necklace?

Wh-questions with an adjunct phrase

7. When do you suppose that he went on a holiday?
8. What do you think he played music with?
9. Where do you presume that she went shopping?

10. When do you suppose she picked up her driver's license?

11. Why do you doubt that she thought of a present?

12. When do you expect he will have made the decision?

Direct object extraction without islands

13. What do you think Paul ate?

14. Which event do you think Robert attended?

15. What do you think Lily planned?

16. Which flower do you think Luke picked?

Medium Fillers (12 items)

Semantically strange imperfective predicate

17. Do you doubt that Elizabeth slept three times last night?

18. Do you think Jack went on an adventure often this morning?

19. Do you suspect he worked twice today?

Semantically constrained passives

20. Do you think many fun times were had by us?

21. How many people are fit by this tent?

22. Who do you think this road was walked by?

Partial wh-movement

23. What do you think which slice covered the last piece of pizza?

24. What do you expect who gave the present?

25. What do you think who sent the postcard?

Wh-copying

26. Who do you suspect who stole the watch?

27. Who do you expect who tasted the wine?

28. Who do you doubt who stole the strawberries?

Bad Fillers (12 items)

Subject islands

29. Who was the story of boring?

30. What hangs a picture of on the wall?

31. Who was the wallet of missing?

32. Which country has the holiday friend from the most chance of a visit from Jill?

Agreement mistakes

33. Do you expect that he have done the groceries?

34. Do you suspect that he have caused the destruction?

35. Does you doubt that she prepared a movie?

36. Does you expect that she selected the laptop?

Word salad

37. Doubt the pike that you caught he have?

38. Suppose the shell that you picked up she have?

39. Expect the white beer that you drink he have?

40. Predict the Siamese cat that you adopt she have?

Appendix 2

Participant Data

In this appendix, detailed data is reported about the participants of the study.

Note that percentages might not add up to 100%, this is due to rounding.

Participant characteristics (n=62)	Data
Native Language (number of participants)	
English	62 (100%)
Gender (number of participants)	
Female	30 (48.4%)
Male	30 (48.4%)
Non-binary	2 (3.2%)
Age (years)	
Mean	23.9
SD	4.7
Range	18-44
Country raised (number of participants)	
United Kingdom	29 (46.8%)
United States	25 (40.3%)
Canada	5 (8.1%)

Participant characteristics (n=62)	Data
Australia	1 (1.6%)
New Zealand	1 (1.6%)
Nigeria	1 (1.6%)
Educational Level (number of participants)	
Did not finish secondary school/high school	3 (4.8%)
Graduated secondary school/high school	20 (32.3%)
Vocational degree/community college	12 (19.4%)
University/college bachelor’s degree	18 (29.0%)
University/college master’s degree	9 (14.6%%)