Investigating Variation in Island Effects: A Case Study of Norwegian Wh-Extraction

Dave Kush
NTNU: Norwegian University of Science and Technology
Haskins Laboratories

Terje Lohndal
NTNU: Norwegian University of Science and Technology
UiT: The Arctic University of Norway

Jon Sprouse
University of Connecticut

Corresponding Author:
Dave Kush
NTNU Norwegian University of Science and Technology
Department of Language and Literature
NO-7491 Trondheim, Norway
dave.kush@ntnu.no

Abstract
We present a series of large-scale formal acceptability judgment studies that explored Norwegian island phenomena in order to follow up on previous observations that speakers of Mainland Scandinavian languages like Norwegian accept violations of certain island constraints that are unacceptable in most languages cross-linguistically. We tested the acceptability of wh-extraction from five island types: whether-, complex NP, subject, adjunct, and relative clause (RC) islands. We found clear evidence of subject and adjunct island effects on wh-extraction. We failed to find evidence that Norwegians accept wh-extraction out of complex NPs and RCs. Our participants judged wh-extraction from complex NPs and RCs to be just as unacceptable as subject and adjunct island violations. The pattern of effects in Norwegian paralleled island effects that recent experimental work has documented in other languages like English and Italian (Sprouse et al. 2012, Sprouse et al. 2016). Norwegian judgments consistently differed from prior findings for one island type: whether-islands. Our results reveal that Norwegians exhibit significant inter-individual variation in their sensitivity to whether-island effects, with many participants exhibiting no sensitivity to whether-island violations whatsoever. We discuss the implications of our findings for universalist approaches to island constraints. We also suggest ways of reconciling our results with previous observations, and offer a systematic experimental framework in which future research can investigate factors that govern apparent island insensitivity.

1. Introduction

Natural languages can establish relations between elements across a distance, a capacity perhaps best exemplified by *filler-gap dependencies* such as *wh-question formation*. In *wh*-questions a 'moved' *wh*-phrase (*which tacos* in 1) is linked to a later *gap* position where it is interpreted (denoted below with an underscore). As shown in (2), these dependencies can be of unbounded length (Chomsky 1973, 1977): a *wh*-word can, in principle, be related to a gap position across a potentially arbitrary linear and structural distance.

(1) Which tacos$_i$ did Sigrid say that Johnny should make ___$_i$?
(2) Which tacos$_i$ did Sigrid say that Torgeir thought that Roar believed that Johnny should make ___$_i$?

Filler-gap dependencies are formally unbounded, but they seem to be constrained. One of the most surprising findings in the history of syntactic theorizing was that fillers cannot be related to gaps inside specific syntactic domains. Ross (1967) christened domains that block filler-gap dependencies 'islands'. A number of constituent types have been identified as islands, including embedded (polar) questions (*whether-islands*), clausal complements of nouns (*Complex NP islands*), complex subjects (*subject islands*), adjunct clauses of various types, and relative clauses (*RC-islands*).

(3) a. *whether-island*
*What do you wonder [whether Sigrid made ___]?
b. *complex NP island*
*What did you make the claim [that Sigrid made ___]?
c. *subject island*
*What did you think that [the recipe for ___] was sitting on the counter?
d. *adjunct island*
*What would you worry [if Sigrid made ___]?
e. *relative clause island*
*What did you meet the woman [who made ___]?

Many theoreticians reason that the data required to determine that such constituents are islands is either extremely rare or non-existent in the primary linguistic data, therefore the existence of islands represents a classic learnability puzzle (e.g., Chomsky 1964, 1973; Lasnik & Saito 1992, Manzini 1992; Phillips, 2013a, b). To solve the learnability puzzle, it is commonly supposed that the unacceptability of the sentences in (3) reflects innate, universal constraint(s) on structure-building. This *Universalist* approach to island phenomena predicts cross-linguistic uniformity with respect to island constraint sensitivity: extraction from embedded questions, RCs, and other islands should have a negative effect on the acceptability of a structure in any language tested. By and large, this prediction has been borne out across a number of different languages (Boeckx 2008, Phillips 2013a, b). Although exceptions have been noted in some languages (Erteshik-Shir 1973, Rudin 1988, Cole & Hermon 1994), it has turned out that many of these exceptions can be given explanations within the Universalist framework (Huang 1982, Richards 2001, Han & Kim 2004, Hoshi 2004, Ishizuka 2009).

There appear to be, however, some recalcitrant exceptions. Most notably, Mainland Scandinavian (MSc) languages, such as Norwegian, Swedish, and Danish, have been reported to

allow filler-gap dependencies across embedded questions (both polar, 4a, and otherwise, 4b) relative clauses (RC-islands), and complex noun phrases (complex NP islands). We focus on Norwegian in this paper, so the examples given below are Norwegian alone, but it is typically implied in the literature that Swedish and Danish counterparts of the examples below would also be judged acceptable.

(4) *embedded questions*

a. Hvem vet du ikke om Jon så *t* på kino?
Who know you NEG whether/if John saw at cinema
'Who don't you know whether Jon saw at the movies?' (Maling & Zaenen, 1982 #3)

b. Hvilke bøker spurte Jon hvem som[1] hadde skrevet *t*?
Which books asked Jon who C had wrote/written
`Which books did Jon ask who had written?' (Maling & Zaenen, 1982 #2)

(5) *RC-islands*

De blomstene kjenner jeg en mann som selger *t*.
Those flowers know I a man who sells
`Those flowers, I know a guy who sells.' (Maling & Zaenen, 1982 #4)

(6) *complex NP islands*

Hvilket fengsel$_i$, er det lite håp om [at man kommer helskinnet fra t_i]?
Which prison is it little hope about that one comes unscathed from
(Maling & Zaenen, 1982 #8b)

Some authors have used the data in (4)-(6) to argue that sensitivity to certain constraints can vary parametrically, or against the idea of universal island constraints altogether (Allwood, 1982; Andersson, 1982; Engdahl, 1982 et seq, Hofmeister & Sag 2010.). The possibility of variation within islands presents a challenge to our understanding of syntactic primitives (and perhaps even analytical proposals that presuppose the universality of islands as constraints on movement). It also has the potential to raise learnability issues thought to be addressed by the Universalist stance: If island constraints are not Universal, how can we explain the consistency in cross-linguistic judgments in the absence of unambiguous primary linguistic input? Or, if constraint sensitivity is subject to parametric variation, what properties of the input trigger parameter-setting? Given the potentially deep implications of these counter-examples, it is important to understand them more fully. To this end, we use the tools of experimental syntax to address three inter-related issues surrounding these MSc island violations.

Our first goal in this paper is to begin the construction of a comprehensive quantitative record of island phenomena in Norwegian, and how judgments of island violations in Norwegian depart from judgments in languages like English. This is an important preliminary step in assessing the limits of cross-linguistic variation since prior research has been based largely on informal acceptability judgments that do not always appear to present a consistent map of the empirical landscape. If there are inconsistencies, quantitative experimental methods can help to

[1] The head *som* obligatorily follows the moved *wh*-phrase in embedded subject questions, but is blocked in embedded questions where a non-subject has been moved. For convenience, we gloss the element as a C head, following Taraldsen (1986), though it has also been treated as an expletive or resumptive that occupies the base position of the subject.

reveal potential patterns in (and causal mechanisms influencing) those inconsistencies. Recent experimental work has begun to provide quantitative information for English, Japanese, and Italian (Sprouse et al. 2011, Sprouse et al. 2012, Sprouse et al. 2016). Applying these methods to Norwegian, a language that has been critical to theories of cross-linguistic variation, further adds to the growing, quantitative, empirical landscape.

Second, we wish to determine whether the acceptability of sentences like those in (5-8) truly reflects the absence of a syntactic constraint violation. It is relatively common in the informal acceptability judgment literature for syntacticians to assume a transparent mapping between the acceptability and the existence of a grammatical constraint violation: if a sentence has relatively high acceptability, there is no constraint violation present, if a sentence is relatively low in acceptability, there is a constraint violation present (and middle levels of acceptability lead to debate in the literature). Quantitative judgment methods allow us to move beyond this mapping and ask instead whether there is an acceptability *effect* present (a difference in acceptability between two or more conditions), regardless of where on the scale this difference occurs. Featherston (2005) famously leveraged this approach to demonstrate that German speakers report the same *pattern* of acceptability for Superiority violations as speakers of English. This finding was particularly surprising given that several (informal) German acceptability studies had previously reported Superiority violations as "acceptable", whereas several (informal) English studies had reported Superiority violations as "unacceptable", leading many researchers to conclude that German lacks whatever constraints give rise to Superiority effects in English (Grewendorf 1988; Müller 1991; Haider 1993; Lutz 1996; Fanselow 2001). By using quantitatively defined effects rather than simple categorical mappings between acceptable/unacceptable and grammatical/ungrammatical, Featherston was able to show that the Superiority *effects* were nonetheless present in German. This, of course, raises many questions about how to interpret the presence of apparent grammatical effects in the absence of "unacceptability" that touch on the very nature of the grammar (see especially Featherston 2005 and Keller 2000 for gradient approaches to grammar). But for our purposes, it raises an interesting question for island effects in Norwegian: Despite the reported lack of categorical unacceptability, are there nevertheless island effects?

Finally, we also wanted to use experimental methods to address a puzzle that has persisted among previous analyses: the source of inconsistency in island judgments. Even though the acceptability of (4-6) is not in dispute, it is not uncommon for speakers to reject seemingly similar island violations. For example, Taraldsen (1982: 206) noted that certain extractions from RCs such as (7) are unacceptable, despite resembling other acceptable cases in many regards (see also Christensen 1982, Allwood 1982; Engdahl 1997, Platzack 2000, Christensen & Nyvad, 2014). More recently, Christensen, Kizach & Nyvad (2013) found that Danish participants gave relatively low ratings to *wh*-island violations in a series of acceptability judgment studies.

(7)	*Rødsprit$_i$	slipper	vi	ingen	som	har	drukket __$_i$	inn
	red.spirit	let	we	nobody	that	has	drunk	in

(Taraldsen, 1982, #9)

On the assumption that acceptable and unacceptable sentences do not differ in their syntactic analysis, theorists have advocated two separate approaches to explaining this unacceptability. One line of reasoning holds that extra-grammatical processing costs are to blame. For example, Christensen and colleagues argued that the relative unacceptability of whether-island violations

in Danish was due to demands that parsing whether-island violations places on individuals' working memory (Christensen, Kazach & Nyvad, 2013, see also Christensen & Nyvad, 2014). Decrements in acceptability associated with processing an island violation are supposed to represent an extreme case of costs associated with processing complex, but otherwise grammatical, sentences (see also Deane, 1991, Kluender & Kutas, 1993, Hofmeister & Sag, 2010 for elaboration of this kind of 'reductionist' view of certain island violations and Sprouse et al. 2012, Phillips 2013a for critical commentary). A second research tradition argues that semantics or discourse-pragmatic conditions are responsible (Erteschik-Shir, 1973; Engdahl 1997). The intuition behind these proposals is that movement dependencies are only acceptable if the transformation is 'motivated' within a discourse. In order to make a judgment, a speaker/hearer must be able to imagine a context in which the sentence would be a felicitous conversational move. Dependencies that span island boundaries impose very stringent demands on their licensing context that are difficult to accommodate when making a judgment *in vacuo*. We make a step toward disentangling these possible sources of unacceptability. As we elaborate below, our experiments employ a design that allows us to factor out linearly additive effects of processing difficulty, which reduces the space of possible explanations to purely grammatical accounts, semantic accounts, or complex (non-linear) processing accounts.

The paper is organized as follows. Section 2 presents the design used across all of our experiments. Section 3 presents the experiments and their results. In section 4, we discuss how the meta-theoretical implications of our settings and discuss ways in which our results could be accommodated within specific theoretical frameworks. Section 5 concludes the paper.

2. A Factorial Design for island effects

Sprouse (2007) developed a factorial design for isolating, and quantifying, *island effects* independently of categorical notions of "acceptable" and "unacceptable" (see also Sprouse et al. 2011, Sprouse et al. 2012, and Sprouse et al. 2016). The factorial design for island effects is typically a 2×2 design, illustrated with an example *whether*-island item in (8). The design crosses two factors, which we label STRUCTURE and DISTANCE, each having two levels. STRUCTURE manipulates the presence of an island configuration. *Non-Island* conditions lack an island, *Island* conditions contain one. Concretely, STRUCTURE determines whether the embedded clause in (8) is an embedded declarative clause (*Non-Island*) or an embedded *whether* question (*Island*). The factor that we label DISTANCE determines the base position of a displaced *wh*-phrase (*who*/*what* in 8). In *Short* conditions the base position of the *wh*-phrase falls in the matrix clause. In *Long* conditions the base position of the *wh*-phrase is located in a more deeply embedded constituent (the embedded CP in 8).

(8) A factorial design for measuring island effects: STRUCTURE × DISTANCE

a.	*Who* __ thinks [that John bought a car]?	NON-ISLAND | SHORT
b.	*What* do you think [that John bought __]?	NON-ISLAND | LONG
c.	*Who* __ wonders [whether John bought a car]?	ISLAND | SHORT
d.	*What* do you wonder [whether John bought __]?	ISLAND | LONG

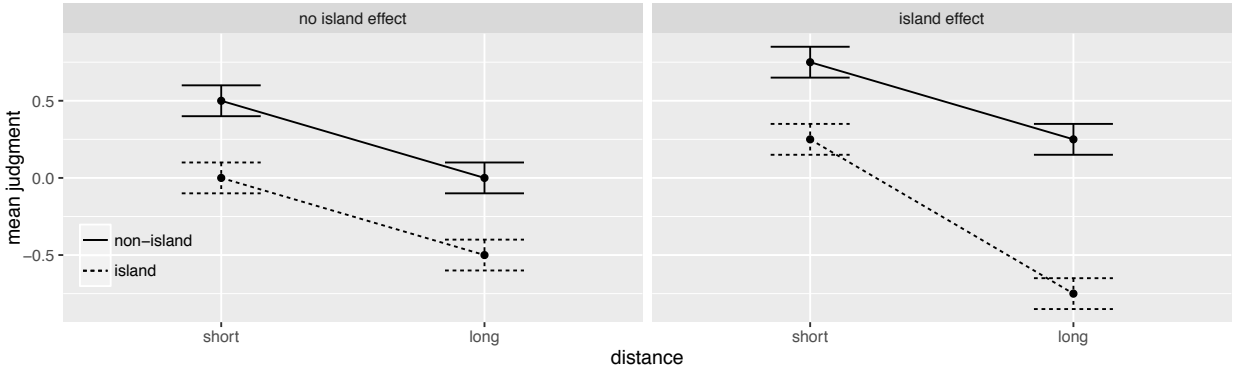
These four sentences allow us to use subtraction logic to do three things. First, we can quantify the difference in baseline acceptability between short (subject) extraction and long-distance (object) extraction as the difference [8a - 8b].
Second, we can quantify the independent acceptability cost of simply having the island structure in a sentence as the difference [8a - 8c]. Finally, we can quantify the remaining effect after those two orthogonal effects on acceptability have been accounted for, which we take to be the *island effect* itself. Mathematically, we can calculate the island effect in two different ways. The first is a simple effects calculation, isolating the two processing costs independently, and the second is called a *differences-in-differences* score, or DD score (Maxwell and Delaney 2003). Both are algebraically equivalent:

island effect = (8a - 8d) - (8a - 8b) - (8a - 8c)
-or-
island effect = (8b - 8d) - (8a - 8c)

One welcome consequence of factorial subtraction logic is that it allows us to potentially control for an unlimited number of confounds, as long as the confounds are distributed across the subtractions such that they subtract to zero in the equations above. For example, (8a) and (8b) are not strict minimal pairs that permit us to perfectly isolate the effect of linear/structural distance. The DISTANCE manipulation is potentially confounded by differences in acceptability between subject and object gap positions, so an observed effect of DISTANCE could also reflect, in part, a difference in the acceptability of subject extraction v. object extraction. This, while true, does not invalidate the subtractive logic for isolating island effects qua interactions (which are actually the measure of interest in this study, as we assume that main effects are not dictated by the grammar). The factorial subtraction logic ensures that we can simultaneously factor out the sum of the individual contributions of each of the dimensions along which the two conditions vary. Similar reasoning applies to the difference between the verbs *think* and *wonder*, and any number of other differences that might arise in these designs. The bottom line is that a two-factor design like this can only *quantify* three effects (the two main effects and the interaction), but it can *control* an unlimited number of potential confounds with respect to the interaction term, as long as the potential confounds are distributed across the subtractions in the correct way.

Factorial designs can be interpreted both visually and statistically. Visually, the factorial design allows us to identify the presence or absence of an island effect by the pattern of the acceptability of the four sentences. If the four sentences, when arranged according to their factors, form two parallel lines, there is no island effect left after the two processing costs have been subtracted. If the lines are not parallel, then there is an island effect present over and above the two processing costs. Figure 1 demonstrates this.

Figure 1: The left panel demonstrates the pattern predicted when no island effect is present. The right panel demonstrates the pattern predicted when there is an island effect.



Statistically, the factorial design allows us to use standard 2×2 analysis techniques (in this case, linear mixed effects models) to identify island effects. The island effect will appear as an interaction term between the two factors (STRUCTURE × GAP-POSITION).

The factorial design has a number of advantages that are relevant for the current study. First, the factorial design quantitatively encodes the intuitive definition of island effects that already exists in the literature. It reveals the effect of a constraint over-and-above the effects of long-distance extraction and island structures. Second, the factorial definition avoids the added layer of complexity imposed by categorical mappings between grammaticality and acceptability, potentially revealing previously unseen effects (similar to Featherston 2005). Third, it quantifies two distinct sources of processing complexity, so we can see exactly how much of an effect long-distance dependencies have on acceptability, and how much of an effect island structures have on acceptability. This allows us to compare simple, linearly additive, processing explanations in which the two processing costs completely explain the unacceptability of island sentences (the left panel of Figure 1) versus complex processing explanations in which the two processing costs must interact to cause the unacceptability (the right panel of Figure 2). In the latter case, it also gives us information about that interaction. For example, if the individual processing effects are relatively small (as is typically the case in English, see Sprouse et al. 2012), then the explanation for island effects must invoke a very large interaction component of some sort. Finally, the factorial design allows us to control for a potentially unlimited number of confounds. Because the factorial design relies on two subtraction steps to isolate the island effect, as long as confounds are placed in the same location in both subtraction steps, they will subtract completely from the isolation of the island effect. This final point merits some elaboration in light of objections of an anonymous reviewer, who correctly noted that conditions in our subtractive designs are not always strict minimal pairs. The reviewer objects that differences above and beyond pure distance could affect acceptability independent of the effect of dependency length and thus confound our ability to quantify the island effect. For example, the *short-* and *long-distance* conditions (8a and 8b, respectively) differ not only in the structural distance between the fronted *wh*-word and its gap, but also in the grammatical role of the extracted phrase. The phrase is a subject in the former, but a direct object in the latter. We agree that the tightest conceivable design would have held grammatical role constant across conditions[2], but offer two comments. First, although the factor is named DISTANCE, it actually quantifies the *aggregate impact of all of the differences* between short- and long-distance conditions, not just dependency-length alone. As long as the residual differences do not interact with STRUCTURE in one of the conditions alone, the contribution of these differences will be subtracted out and will not impede our ability to quantify the island effect. To be concrete: we have no reason to expect the effect of subject vs. object extraction to affect *island* conditions more adversely than the *non-island* conditions, so we consider the subtractive logic suitable for controlling for any acceptability differences that emerge as a result of this choice. If the goal of the study had been to isolate the independent cost of linear distance (or some other effect), then stricter minimal pairs, or a design that triangulated the contribution of the specific aspect through additional comparisons, would have been warranted. However, since our primary aim was to isolate the interaction effect itself, not the various factors that contributed to main effects between conditions, we do not view the absence of strict minimal pairs to be detrimental.

[2] One possibility – for many, but not all of our experiments - would be to extract indirect objects from matrix and embedded clauses. We encourage future researchers to conduct this comparison if they are concerned about the potentially confounding effect of grammatical role discrepancies.

Second, there are often practical challenges associated with using strict minimal pairs, which would introduce troublesome interactions. For example, choosing to extract subject DPs across all four conditions would have led to complementizer-trace configurations in the island violation condition, which would have vitiated our measurement of the island violation alone. We endeavored to construct our materials in such a way as to avoid introducing differences that would interact with long-distance extraction in the island-violating condition. The same kind of reasoning applies to other differences – such as differences in embedding predicates – across conditions.

3. Experiments

We ran three acceptability judgment studies that tested island sensitivity in Norwegian. The three studies were very similar in design, analysis, and outcome. Thus, we report the results of all three studies at once in the interest of space. We note wherever studies differed in their procedure or design.

All three experiments that we conducted tested minimally tested four island types in Norwegian: whether-islands, Complex NP-islands, subject islands, and (conditional) adjunct islands. We chose these islands for two reasons. First, they enable a direct comparison with the results of previous experiments that have used the factorial design to test the same island types in English (Sprouse, 2007; Sprouse et al., 2012), Japanese (Sprouse et al., 2011), and Italian (Sprouse et al., 2016). Second, these four island effects are the "better" versions of four structurally-similar island types (cf. *wh*-islands, Relative Clause islands, Sentential subject islands, and causal adjunct islands). They were originally chosen in the previous studies because they are (anecdotally) reported to lead to relatively higher acceptability ratings, and reported to lead to more variability among speakers. These continue to be desirable properties for the current study. In addition to the four island types mentioned above, experiments 2 and 3 also tested Relative Clause islands.

Experiment 1 tested the acceptability of extracting a bare *wh*-word (e.g., 'what') from *whether*, complex NP, subject, and adjunct island configurations in Norwegian. Our goal was to establish quantitative baselines for these four islands in Norwegian that could be directly and quantitatively compared to the English results of Sprouse et al., (2012). We chose not to test RC islands in experiment 1, despite the fact that they have occupied a central position in many discussions of island effects in Scandinavian, because we wanted to maximize the similarity between our experiment and Sprouse et al., (2012).

Experiment 2 had two goals. First, we sought to test whether the results of Experiment 1 would replicate. Second, we wished to extend the factorial design to investigate relative clause islands given the focus on relative clause islands in the theoretical literature.

Experiment 3 tested whether the pattern of island effects differs when the extracted element is a complex *wh*-phrase (e.g., 'which tacos'), either in terms of the aggregate super-additive interaction or in terms of the individual variation in the interaction.

Our motivation for testing the effect of filler-complexity on island effects was two-fold. First, the majority of attested examples of acceptable island violations involve the movement of a complex, rather than bare, filler. Second, it has been suggested that complex *wh*-phrases may ameliorate island effects, with recent quantitative investigations in English yielding conflicting results (Goodall 2015, Sprouse et al. 2016).

3.1 Materials

Materials for the first four island types were adapted translations of the English items used in Sprouse et al., (2012). For each island type, eight sets of test sentences were generated. Bare *wh*-fillers (e.g., *hvem* 'who') were used in experiments 1 and 2, whereas complex *wh*-fillers (e.g., *hvilken gjest* 'which guest') were used in experiment 3. An example set from the whether-island experiment is in (9).

(9) **whether-island**

a. {Hvem / Hvilken gjest} ___ tror [at Hanne bakte kaken?]
Who / Which guest thinks that Hanne baked cake.DEF
'Who/Which guest thinks that Hanne baked the cake?'

b. {Hva / Hvilken kake} tror gjesten [at Hanne bakte ___?]
What Which cake thinks guest.DEF that Hanne baked
'What/Which cake does the guest think that Hanne baked?'

c. {Hvem / Hvilken gjest} ___ lurer på [om Hanne bakte kaken?]
Who Which guest wonders on if/whether Hanne baked cake.DEF
'Who wonders whether Hanne baked the cake?'

d. {Hva / Hvilken kake} lurer gjesten på [om Hanne bakte ___?]
What Which cake wonders guest.DEF on if/whether Hanne baked
'What does the guest wonder whether Hanne baked?'

There is one potentially noteworthy difference between the *whether*-island configurations in Sprouse et al. (2012) and the current experiment. Unlike *wonder*, the equivalent Norwegian verb *lurer* does not take a CP complement directly. Instead, the CP must be the complement of a preposition *på* 'on', which is selected by the verb (Åfarli & Eide 2003). The logic of the factorial definition of islands allows us to subtract out any main effect of additional structural complexity contributed by the preposition. However, the factorial design does not allow us to factor out a potential effect of A'-movement from out of the prepositional phrase itself in the *long-island* condition (9d). This means that the island effect that we quantify here will be the sum of the effect of extraction from the embedded *whether*-question and the effect of extraction from a prepositional phrase. We do not consider this a serious confound, because we believe that if extraction out of a PP affects acceptability, the effect should be negligible, because prepositions typically do not block long-distance movement that originates within their complements. There are two pieces of evidence for this claim: (i) Norwegian is a preposition-stranding language (15a), and (ii) long-distance movement is allowed, for example, from (declarative) clausal complements of prepositions selected by verbs such as *å insistere* ('to insist'), as in (15b), though we know of no formal experiments that quantify these judgments.

(10) a. Hvem$_i$ snakket regissøren med ___$_i$?
Who spoke director.DEF with
'Who did the director speak with?'

b. Hva insisterte John [på [at mannen måtte lese ___]] ?

What insisted John on that man.def must read
'What did John insist that the man must read.'

An example set for from our complex NP island experiments is in (11).[3]

(11) **complex NP island**

a. {Hvem / Hvilken dommer} ___ rapporterte at Anders vant medaljen?
Who / Which judge reported that Anders won medal.DEF
'Who/Which judge reported that Anders won the medal?'

b. {Hva / Hvilken medalje} rapporterte dommeren at Anders vant ___?
What / Which medal reported judge.DEF that Anders won
'What/Which medal did the judge report that Anders won?'

c. {Hvem / Hvilken dommer} ___ rapporterte nyheten om at Anders vant medaljen?
Who / Which judge reported news.DEF about that Anders won medal.DEF
'Who/Which judge reported the news that Anders won the medal?'

d. {Hva / Hvilken medalje} rapporterte dommeren nyheten om at Anders vant ___?
What / Which medal reported judge.DEF news.DEF about that Anders won
'What/Which medal did the judge report the news that Anders won?'

A subject island set is in (12).

(12) **subject island**

a. {Hvem / Hvilken journalist} ___tror [at møtet forsinket den politiske enigheten?]
Who / Which journalist thinks that meeting.DEF destroyed the political union
' Who/Which journalist thinks that the meeting destroyed the political union?'

b. {Hva / Hvilken møte} tror journalisten [___ forsinket den politiske enigheten?]
What / Which meeting thinks journalist.DEF destroyed the political union
' What/Which meeting does the journalist think destroyed the political union?'

c. {Hvem / Hvilken journalist} __ tror [at møtet med millionæren forsinket den politiske enigheten?]
Who /Which journalist thinks that meeting.DEF with millionaire.DEF destroyed the political union
'Who/Which journalist thinks that the meeting with the millionaire destroyed the political union?'

[3] The observant reader will note that the complement of the noun *nyheten* ('the news') is a PP, headed by *om* ('about'), rather than a bare CP, as in English. Clausal complements to N must always be wrapped in a PP (see Lødrup, 2004), thus this difference from the English examples is unavoidable.

d. {Hvem / Hvilken millionær} tror journalisten [at møtet med ___ forsinket den politiske enigheten?
Who / Which millionaire thinks journalist.DEF that meeting.DEF with destroyed the political union?
' Who/Which millionaire does the journalist think that the meeting with destroyed the political union?'

One potential issue with this design for subject islands (raised by Caroline Heycock, p.c.) is that the effect isolated in the interaction term will contain both the subject island effect, and any potential independent effect of sub-extraction (i.e., an effect of extracting out of a complex NP regardless of its structural position). If the sub-extraction effect exists, it means that the design in (12) will overestimate the size of the subject island effect. We settled on (12) instead of a design that directly controlled for sub-extraction effects (as in 13 below, previously explored by Sprouse 2007 and Sprouse et al. 2011) because (13) has the reverse problem: it would systematically underestimate the size of the subject island effect.

(13) *A subject island design that controls for sub-extraction*
a. What do you think the meeting destroyed __?
b. What do you think __ destroyed the consensus?
c. What do you think [the meeting about the amendment] destroyed [the consensus over __] ?
d. What do you think [the meeting about __] destroyed [the consensus over the proposal]?

The design in (13) underestimates the subject island effect because it has two confounds. First, there is a filled-gap effect at *the consensus* in (13c) that is not balanced out in any other condition. This effect decreases the island effect in the subtraction logic (see Sprouse 2008 for evidence that filled-gap effects lower acceptability even in offline experiments). Second, adding complex NPs in both subject and object position in (13c) and (13d) to control for overall DP/NP complexity substantially lowers the acceptability of these conditions, potentially causing a floor effect that limits the size of the subject island effect (see Sprouse 2007 and Sprouse et al. 2011 for mean ratings of these two conditions in English). This leads to a difficult choice between potentially overestimating the subject island effect (if sub-extraction is an independent effect), or definitely underestimating the subject island effect (because filled-gap effects are established in the judgment literature). We opted for (12) because the effect of sub-extraction has not been independently quantified in the literature to our knowledge, and even if it exists, it is likely to be substantially smaller than an island effect (e.g., nobody has ever claimed that there are "object" island effects in English). We considered it better to risk a slightly over-inflated island size than to risk a null result that is ambiguous between no island effect and a small island effect that is obscured by the confounds. Our results (reported in section 3 below) suggest a relatively large subject island effect in all three experiments. This effect is roughly the same size as subject island effects in English, so we tentatively conclude that it is a (potentially overinflated) subject island effect, and not the (likely smaller) sub-extraction effect alone.[4]

[4] Given that the subject island effects that we observed are relatively large, an experiment that tested the Norwegian equivalent of (13) would serve as an excellent cross-validation. We expect that the subject island effect should be

(14) is an example adjunct island set.

(14) **adjunct island (*if-clause*)**

a. {Hvem / Hvilken person} tror [at advokaten glemte mappen sin på kontoret?]
Who /Which person believes that lawyer.DEF forgot folder.DEF his at office.def
'Which person believes that the lawyer forgot his folder at the office?'

b. {Hva /Hvilken mappe} tror du at advokaten glemte ___ på kontoret?
What Which folder believe you that lawyer.DEF forgot at office.DEF
'What/Which folder do you think that the lawyer forgot at the office?'

c. {Hvem / Hvilken person} ___ er glad om advokaten glemte mappen sin på kontoret?
Who Which person is happy if lawyer.DEF forgot folder.DEF his at office.def
'Who/Which person is glad if the lawyer forgot his folder at the office?'

d. {Hva /Hvilken mappe} er du glad om advokaten glemte på kontoret?
What Which folder are you happy if lawyer.DEF forgot at office.DEF
'What/Which folder are you happy if the lawyer forgot at the office?'

Test items for RC-islands in experiments 2 and 3 used the factorial design illustrated in (15). In order to maximize the likelihood that RC island violations would be judged acceptable, our test items shared were modeled after attested examples of acceptable RC island violations. We used indefinite subject RCs as our test island because attested examples commonly feature subject RCs (Platzack 2000; Engdahl, 1997, Lindahl, 2014) with indefinite or weak quantificational heads (Engdahl 1982, 1997).[5]

(15) **relative clause island**

a. {Hvem / Hvilken regissør} __ trodde at et par kritikere hadde stemt på filmen?
Who / Which director thought that a few critics had voted for film.DEF
'Who/Which director thought that a few critics had voted for the film?'

b. {Hva / Hvilken film} trodde regissøren at et par kritikere hadde stemt på __?
What / Which film thought director.DEF that a few critics had voted for
'What/Which film did the directory think that a few critics had voted for?'

large enough to survive the filled-gap effect and the potential floor effect of the design in (13). We leave such a validation to future research.

[5] Some authors have proposed that only subject RCs allow extraction (Platzack, 2000; Kush, Omaki & Hornstein, 2013), but this has been disputed (Engdahl, 1997; Lindahl, 2014). It was also initially proposed that indefiniteness is a necessary condition for acceptable RC extraction, but this claim is contradicted by some attested examples (Maling & Zaenen, 1982; Engdahl, 1997).

c. {Hvem / Hvilken regissør} __ snakket med et par kritikere som hadde stemt på filmen?
Who / Which director spoke with a few critics that had voted for film.DEF
'Who / Which director spoke with a few critics that had voted for the film?'

d. {Hva / Hvilken film} snakket regissøren med et par kritikere som hadde stemt på __?
What / Which film spoke director.DEF with a few critics that had voted for
'What / Which film did the director speak with a few critics that had voted for?'

Two properties of our RC island test items merit discussion. First, the four conditions were not as closely lexically-matched as in other islands because the matrix verbs differed between island and non-island conditions. In non-island conditions the matrix verb was a propositional attitude verb that embedded a declarative CP complement. In island conditions the matrix verb was either a simple transitive verb (e.g. *møtte* 'met') or a V-P string (e.g. *snakket med* 'spoke with'). An ideal manipulation would have held the embedding verb constant across conditions by using verbs that take both DP and CP complements, but we reasoned that this was not possible. Verbs such as *se* ('see') or *vet* ('know') that take DP and CP complements in Norwegian were considered, but using these verbs would have resulted in an unintended confound: *long/non-island* conditions would have instantiated factive island violations (Rouveret 1980, Kayne 1981, Zubizarreta 1982, Adams 1985).

(16) ?*What did the director see/know that the critic voted for?

We acknowledge that the difference in verb between non-island and island conditions is a minor confound in the quantification of the effect of STRUCTURE. However, we point out that this difference does not confound the quantification of the island effect itself: the two-step subtraction logic of the factorial design eliminates this effect, just as it does with the change of predicates with other island items above.

The second potential issue in the materials is that the DP containing the RC was the complement of a preposition (*med 'with'* above) in seven of eight test items. As with the whether-island items, any main effect that the presence of the preposition has on acceptability is subtracted out by the factorial design, but the effect of the extraction from the prepositional phrase is not. Thus, the interaction effect represents the sum of the actual RC-island effect and the effect of extraction from a prepositional phrase. Once again, we believe that extraction out of PP should not adversely affect acceptability. In support of this, we provide an analysis of each RC-island item in section 3.5 (Figure 3) to demonstrate that there is no difference between the preposition items and the no-preposition item. Therefore we are confident in the ability of these experiments to accurately estimate the size of the RC-island effect.

3.2 Participants

Ninety-eight Norwegian speakers participated in experiment 1 (mean age 32.3, sd=10.4, 51 female). These participants were recruited either through a public post on Facebook, or through an undergraduate class at the Norwegian University of Science and Technology (NTNU).

Participants provided their age and gender, and were asked to report their first language, their dominant language, and any languages that they had significant exposure to as a child. We excluded four of the original ninety-eight participants from further analysis because they failed to identify Norwegian as their native and/or dominant language. Fifty-one different individuals (mean age 29.3, sd= 9.8, 30 female), recruited through the same channels, participated in experiment 2. Five participants were excluded because they failed to identify Norwegian as their native and/or dominant language. Seventy-four new individuals participated (mean age 30.8, sd= 11.0, 42 female) participated in experiment 3. Ten participants were excluded from analysis because they reported that Norwegian was either not their native or first language. Data were excluded from one additional participant who took under 100ms to respond on numerous trials. All participants took part voluntarily.

3.3 Procedure

In all three experiments participants completed a survey hosted on IbexFarm (Drummond, 2012). Each survey contained 2 tokens of each of 4 conditions for each island type in the experiment. In experiment 1, this meant that participants rated 32 test items (2 tokens × 4 conditions × 4 island types), while in experiments 2 and 3 participants rated 40 test items (2 tokens × 4 conditions × 4 island types). Test items were interspersed pseudo-randomly among 48 filler sentences (16 acceptable, 32 unacceptable; 36 declarative, 12 interrogative; leading to a roughly even balance of acceptable to unacceptable sentences and declaratives to interrogatives). Fillers ranged from simple mono-clausal to multi-clausal sentences. Unacceptable sentences contained a variety of violations ranging from basic morpho-syntactic mismatches and word-order violations to subtler semantic and syntactic violations. The complexity and range of filler sentences was varied so as to encourage participants to make use of the full range of the ratings scale. In order to complete the survey, participants read one sentence at a time and were asked to judge its acceptability on a 7-point scale, with 1 labeled *Dårlig* ('bad') and 7 labeled *Bra* ('good').

3.4 Analysis

Raw ratings were z-score transformed by participant in order to eliminate biases in how different participants used the 7-point scale. We analyzed the z-scored ratings using linear mixed effects models with fixed effects of STRUCTURE, DISTANCE and their interaction. We report the results of models with random intercepts for both subject and item and by-subject random slopes for all fixed effects and their interaction. We calculated *p*-values for main effects of STRUCTURE and DISTANCE and the STRUCTURE × DISTANCE interaction term using likelihood ratio tests. Differences-in-differences (DD) scores were first calculated for each participant, and then averaged across participants for each island. This averaging provided a non-standardized effect-size for each island type.

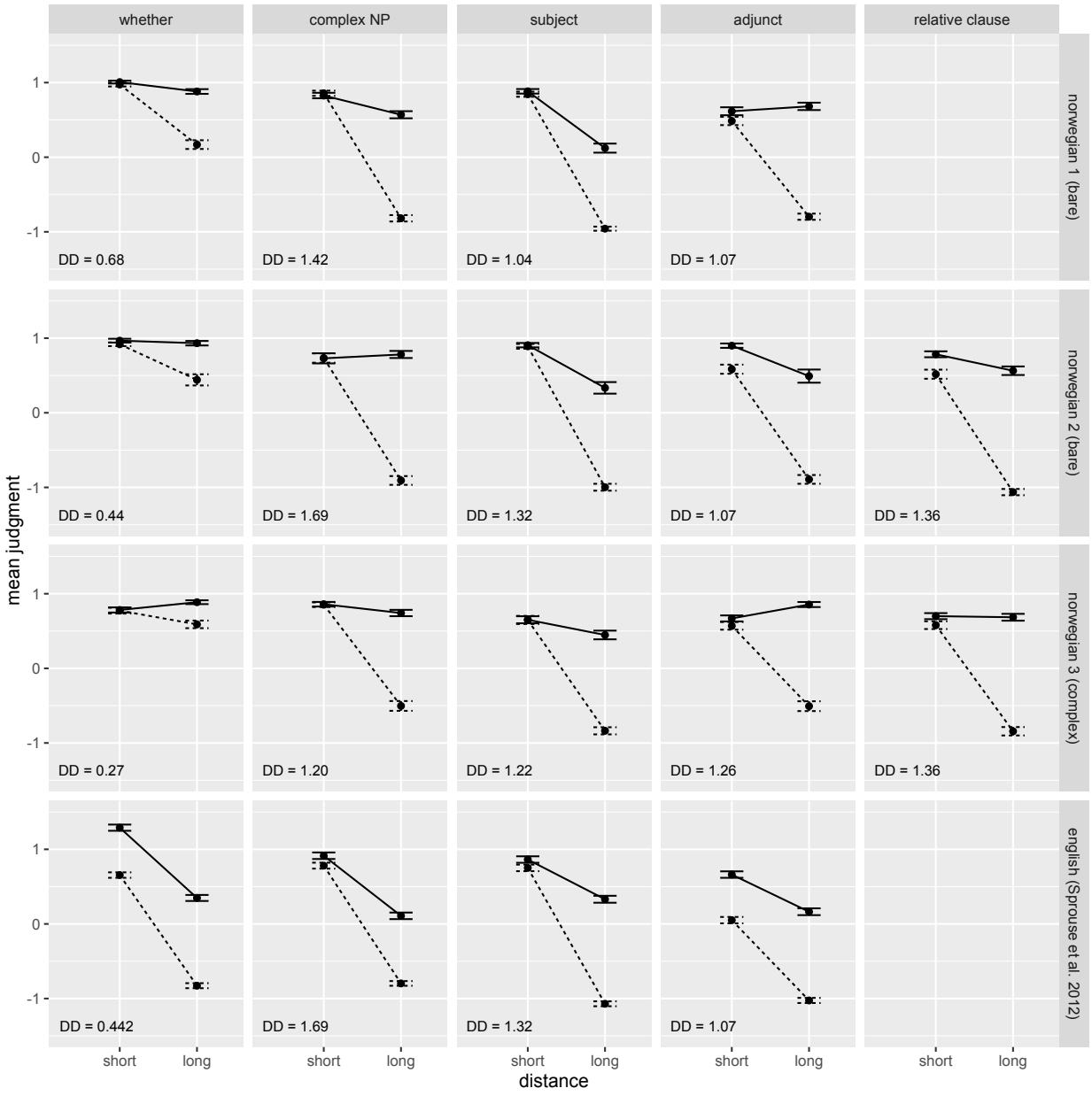
3.5 Results and Discussion

Figure 2 plots the mean ratings for each island type (by column) for each experiment (by row). The fourth row of Figure 2 presents the English results of Sprouse, Wagers & Phillips' (2012) experiment 2 for comparison.

The first result to note is that there appear to be super-additive interactions for all island types that we tested in all three Norwegian experiments. The super-additive effects that we observe all conform to the configuration typical of island effects (cf. Figure 1): island violating sentences receive much lower z-scored ratings than any of the sentences in their paradigm.

Statistical analysis using linear mixed effects models reveals all interaction effects in Figure 2 to be significant at at least the $p < .01$ level. The size of each island effect, measured by DD score, is listed on its respective sub-plot. We discuss each island effect individually.

Figure 2: Interaction plots for all three Norwegian experiments (rows 1 – 3) and the effects from Sprouse et al. (2012) in row 4 for comparison.



Subject island effects were found across all three experiments (all *p*s < .001). The magnitude of the subject island effect, as measured by DD score, was consistently large: all DD scores were greater than one (an effect size that is equal to roughly one standard deviation of the mean given that the ratings were z-score transformed). Norwegian speakers appear to judge subject island-violating sentences as profoundly unacceptable, as demonstrated by the fact that the average z-scores of the island violating sentences cluster around -1. As a point of reference, the average z-scored acceptability ratings of unacceptable filler sentences across all three experiments were near, but slightly greater than -1 (mean rating from experiment 1: -0.78, experiment 2: -0.89, and experiment 3: -0.81).[6] The size of subject island effects in all three Norwegian experiments, as measured by DD score, were comparable to subject island effects in English (Sprouse, Wagers & Phillips, 2011, experiment 2: 1.25).

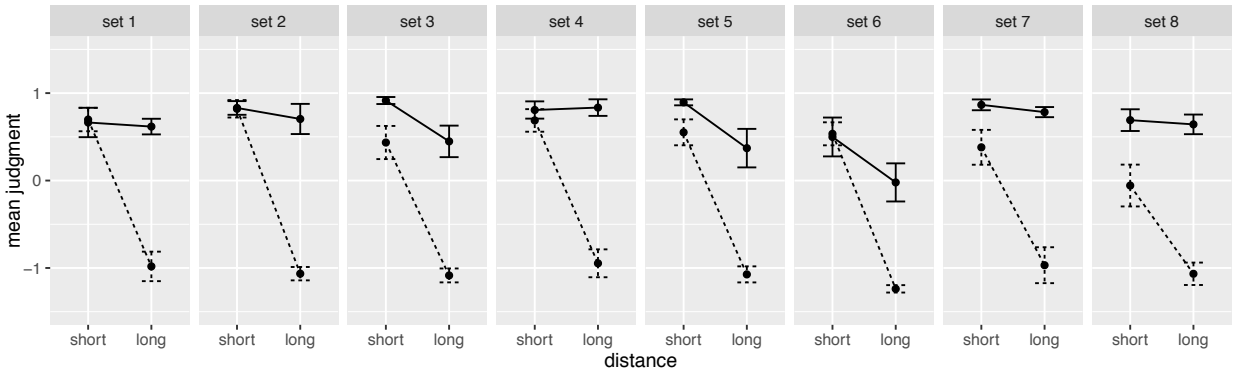
Results of the adjunct island sub-experiments were very similar to the subject island results (all *p*s < .001). Participants assigned very low average ratings to adjunct island violations across all three experiments and the adjunct island effect sizes were consistently above 1. These effects are analogous to adjunct island effects reported in previous experiments for English (Sprouse, Wagers & Phillips, 2012: 1.04, 0.61; Sprouse et al. 2016: 0.71) and Italian (Sprouse et al. 2016: 1.31). In sum, Norwegian judgments of subject and adjunct island effects seem to align very closely with the cross-linguistic norm.

Norwegian judgments of complex NP island violations, perhaps surprisingly, follow the same pattern as observed with subject and adjunct island violations (all *p*s < .001). Participants' average ratings of complex NP violations fell near the low end of the scale, and the average size of the complex NP island effect was comparable to adjunct and subject island effect sizes. Once again, the size of the complex NP island effect falls well within the range of the effects cross–linguistically (English CNP: Sprouse, Wagers & Phillips, 2012: 0.98, 0.80; Sprouse et al. 2016: 1.05; Italian: Sprouse et al. 2016: 0.89).

Judgments of RC island violations in experiments 2 and 3 resemble the judgments of subject, adjunct, and complex NP island violations (all *p*s < .001). The size of the RC island effect was similar to those three island effects and the mean rating of the RC island-violating sentence was as low (or lower) than other island-violating sentences. Given that the data depart from the consensus view that RCs are not islands in MSc languages, we attempted to root out any possible confounds that might have contributed to an illusory RC island effect. As mentioned above, one potential concern with the RC island design is that in seven of eight of the items, the DP containing the RC was complement to a preposition. One might worry that the unacceptability should not be linked to extraction from the RC, but rather to extraction out of the PP. In order to determine whether the preposition was driving the effect, we created interaction plots for each of the 8 sentence-sets for the relative clause island design (Figure 3). All eight show the super-additive pattern, including the item that did not have a preposition (item 6). This suggests that there is a super-additive RC island effect over and above the effect of extracting out of the PP.

[6] An appendix containing all test and filler materials, as well as by-item summary statistics for filler items have been included as Supplementary Materials.

Figure 3: By-item interaction plots for the RC island design. Ratings in all eight sentence sets show an island super-additive effect, including the set 6, which did not involve a preposition.



Our participants' judgments of *whether*-islands differed from their judgments of any other islands that we tested in two related ways. First, although the interactions were significant in all three experiments ($p < .001$, $p < .001$, $p < .01$, respectively), the super-additive whether-island effects across the three experiments were noticeably smaller than other island effects. DD scores of *whether*-island effects were consistently (and significantly) lower than 1. Norwegian *whether*-island effects were also smaller than whether-island effects measured in other languages. Movement of a bare *wh*-word from a *whether*-island in Norwegian led to effects that were roughly half the size (DDs = 0.69, 0.44, in experiment 1 and 2, respectively) of the effects that the same movement produced in English (Sprouse, Wagers & Phillips, 2012: DD = 1.09, 0.87; Sprouse et al. 2011: DD = 1.15), or Italian (Sprouse et al. 2016: DD = 1.69). Extraction of a complex *wh*-phrase in Norwegian also resulted in a much smaller effect (DD = 0.28, experiment 3) than Sprouse et al. (2016) observed in English (DD = 0.62). Second, the average z-scored rating of a *whether*-island violation is above zero in all three experiments. Positive z-scores are typically reserved for sentences whose acceptability is not in dispute: Consider the fact that the ratings of whether island violations are numerically similar to judgments of grammatical *long/non-island* sentences in the subject island sub-experiment (experiments 1 and 2).

Taken at face value, the results might seem to suggest that although there is a *whether*-island effect in Norwegian, violating a *whether*-island has a less severe negative effect on acceptability in Norwegian than it does in English (or Italian). This interpretation would be consistent with Featherston's (2005) claim that syntactic constraints apply in all languages, but that the *strength* of a violation may vary cross-linguistically. While certainly a possibility, we point out that this interpretation is only valid if the aggregate data are representative of a consistent pattern of effects across participants. Under this interpretation, the majority of participants should show a DD score close to the aggregate mean and assign 'intermediate' acceptability ratings to *whether*-island violations. On the other hand, it is also possible that the intermediate results reflect artifacts of an averaging process that obscures a more complex pattern of judgments across participants. In order to tease these two possibilities apart, we examined the individual participant data more closely for signs of variability.

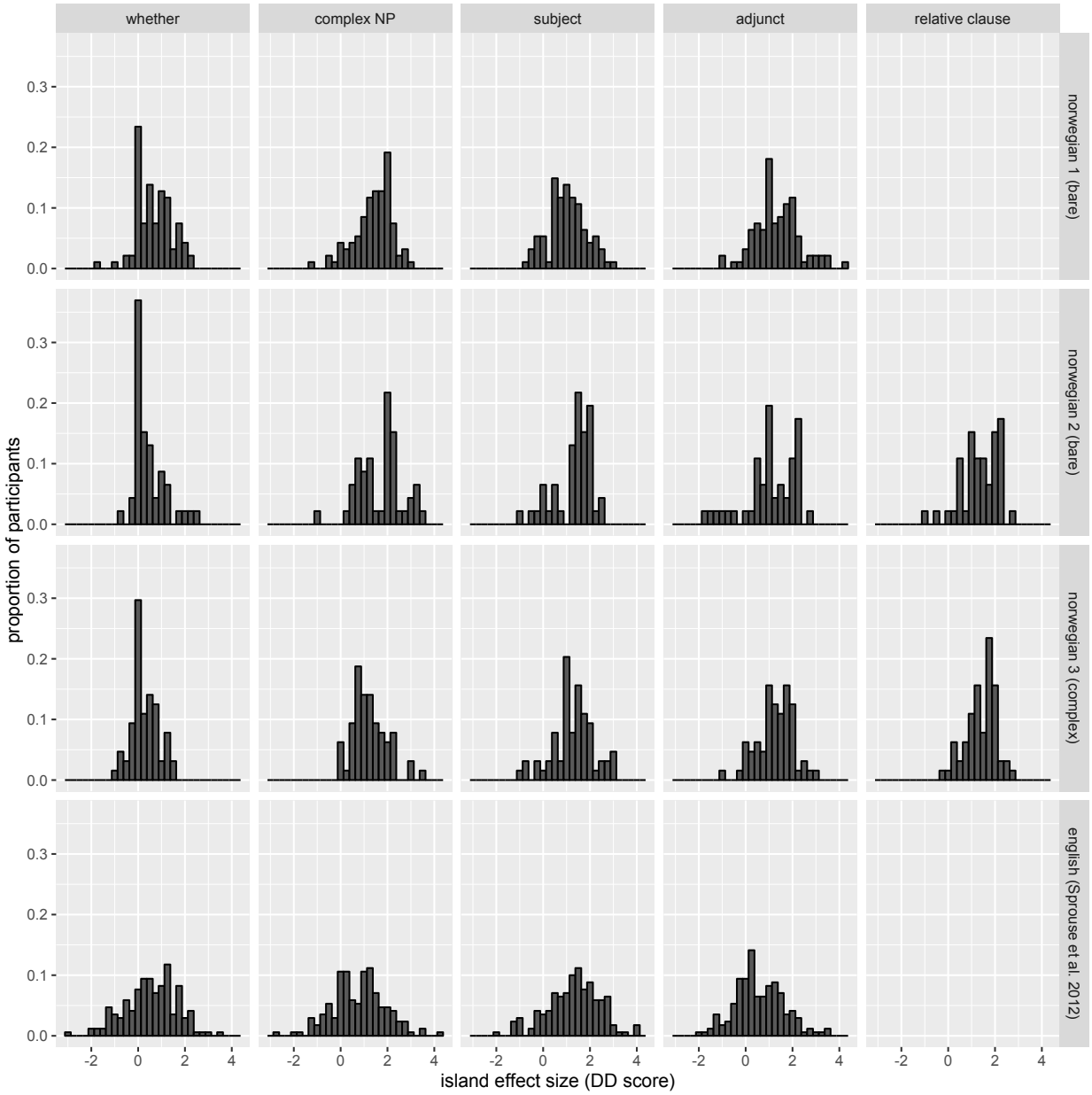
First, we inspected the distributions of individual participants' whether-island DD scores in each experiment and compared them to the distribution of DD scores for other islands. We also compared the distribution of Norwegian DD scores to English island effects from Sprouse et al. (2012), experiment 2. We conducted this comparison to ascertain whether the small whether-island effects in our experiments reflect consistently small DD scores across all participants. Figure 4 plots the distribution of DD scores for islands (by column) and experiments (by row).

Figure 4 reveals important differences between Norwegian *whether*-islands on the one hand and the rest of the islands on the other. The distributions of Norwegian complex NP, adjunct, subject, and RC island effects are roughly (i) unimodal, and (ii) symmetrically distributed about the observed mean DD score. The distributions of island effects in Sprouse and colleagues' English data follow a similar pattern. These distributions reflect a high degree of consistency across participants for each of these islands. The distributions of *whether*-island effects in the Norwegian experiments follow a different pattern. Most notably, we see that a large number of participants in all three experiments had DD scores within between 0 and 0.25: nearly 30% of participants in experiment 1, 52% of participants in experiment 2, and 47% of

participants in experiment 3. This suggests that a significant portion of Norwegian participants in each experiment showed absolutely no *whether*-island sensitivity whatsoever.[7]

[7] We found no consistent age, gender, or dialect differences between groups of accepters and rejecters.

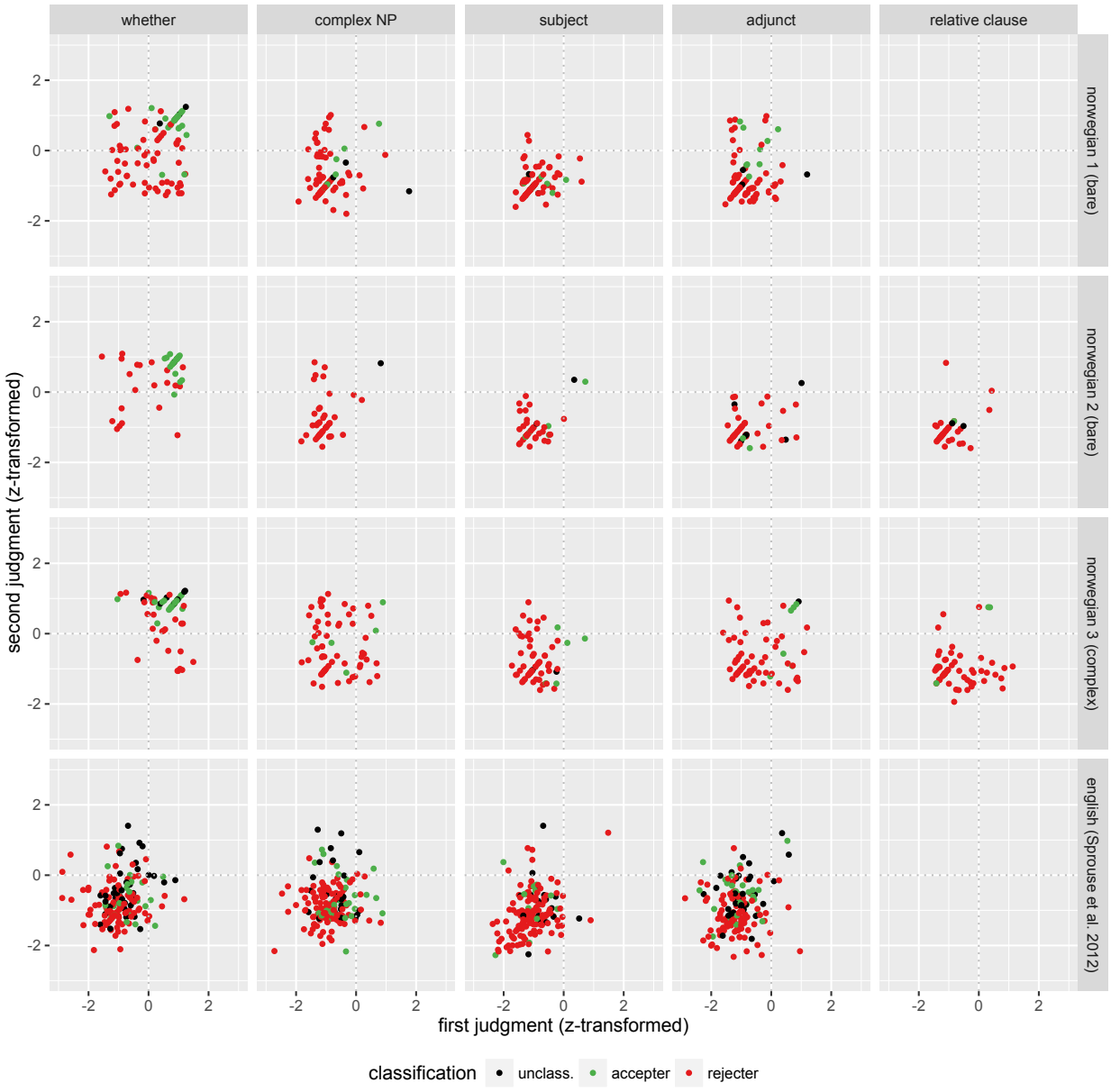
Figure 4: The distribution of island effect sizes (DD scores) by participant for Norwegian experiments 1 – 3 and English (row 4). English data are taken from the first two judgments per island type from Sprouse et al. 2012, experiment 2.



The distributions in Figure 4 suggest that the intermediate DD scores that we observed at the population level are not due to consistently smaller individual island effects. Instead, they seem to arise from averaging across data that is characterized by substantial inter-speaker variability.

The analysis of DD scores above does not provide direct insight into the source of the intermediate average acceptability rating of the *whether*-island violating sentences (because it is an analysis of all 4 conditions simultaneously). As discussed above, we wanted to determine whether participants consistently rated *whether*-island violations around the midpoint of the scale or whether the appearance of relative average acceptability was caused by averaging over ratings that displayed a degree of variability similar to the one we saw with DD scores. To this end we created scatterplots to show how consistently participants were in their rating of *whether*-island violations across the two tokens that they saw during the experiment. The scatterplots in Figure 5 show the relationship between each individual's first and second rating of the island-violating condition (long/island) for every island type tested. In each plot a single dot represents an individual participant. Major axis lines through the 0s (the mid-point of the z-transformed rating scale) divide each plot into four quadrants. Participants that fall into quadrant 1 (upper right) are those participants who consistently rated island-violating sentences above 0 (the mid-point of the mean z-score scale). Quadrant 3 (lower left) indicates participants who consistently rated island violating sentences below 0. The other two quadrants (2 and 4) indicate inconsistency: one highly rated token and one low rated token. Given the aggregate results and the consistency in the size of complex NP, adjunct, subject, and RC island effects, we expected that most participants would fall into quadrant 4 for these strong islands. We were interested in seeing if participant judgments of whether-island violations would pattern differently: if participants consistently rated island violations at the midpoint of the scale we would expect the majority of participants to cluster around the origin in *whether*-islands. Otherwise, we would expect a more diffuse distribution of participants throughout the ratings space. Before moving on to discuss the scatterplots in detail, we note that the colors of the dots on our plots indicate whether the participant showed an island effect (or not) in their DD scores. For concreteness, we defined three categories of participants: island *rejecters* had a DD score above .25, island *accepters* had a DD score between -.25 and .25, and unclassified participants have a DD score below -.25 (a pattern that is not interpretable given current theories). This coloring scheme aids in identifying whether participants who showed no island effect also consistently accepted *whether*-island violations. Doing so is important because, although a DD score of close to zero indicates that a participant showed no island effect, it does not guarantee that that participant actually *accepted* island violations.

Figure 5: The ratings of island-violating sentences, per participant, for Norwegian experiments 1 - 3 and English (Sprouse et al. 2012, experiment 2, the first two tokens) for comparison. The × and y-axes show the rating of the first and second tokens, respectively, per participant. Each dot corresponds to a participant and dot color indicates whether the participant is an island rejecter (DD > .25), island accepter (DD within .25 of 0), or unclassified (DD < -.25).



Consistent with our predictions, the scatterplots reveal that Norwegian participants judged complex NP, subject, and adjunct island violating sentences to be unacceptable with relative consistency. The vast majority of participants' ratings occupy quadrant 3, and very few are found in quadrant 1. Ratings for these three islands in Norwegian also align very closely with the English judgments from Sprouse et al (2012), plotted on the fourth row of Figure 5.

Judgments of *whether*-islands in Norwegian once again show a markedly different pattern from other Norwegian islands and all English islands. *Whether*-island judgments displayed an unexpected amount of variability both across and within participants. There was also a fair amount of variation in judgments across experiments. Ratings in experiment 1 were distributed across all four quadrants, though quadrant 4 had the fewest participants. Many participants fell into quadrant 1, indicating that they rated both whether-island violations above 0. Twenty-seven of the participants in quadrant 1 were accepters (green dots) according to our classification scheme based on DD score. We can be confident that these participants are 'true accepters', that is, their consistently high ratings and negligible DD scores together indicate that they find *whether*-island violations unobjectionable. Slightly fewer participants rated both tokens below 0 in experiment 1. The remainder of participants were inconsistent, tending to accept the first whether-island violation that they rated and reject the second. The existence of inconsistent raters is somewhat mysterious, as it is not immediately clear how to accommodate inconsistent ratings in current theories. It is possible that the inconsistent ratings simply reflect experimental noise, however we find this explanation implausible given how many participants fall into this category. In experiment 2, most participants fell into quadrant 1. Twenty-three of these participants were 'true accepters', with DD scores of approximately 0. Far fewer participants rejected *whether*-island violations (either consistently or inconsistently) in experiment 2 than in experiment 1. We do not have an explanation for this difference, though we speculate that it could be partly due to the difference in sample sizes between experiments or differences between the sample populations (the sample population for experiment 1 consisted of students and non-students from a wider age range, though we found little correlation between age and DD score in post-hoc analyses). Ratings of *whether*-island violations in experiment 3 were very similar to ratings in experiment 2: The overwhelming majority of participants rated both *whether*-island tokens above 0 (nearly half of which were consistent accepters), while there were a few inconsistent raters. Only one participant consistently rejected *whether*-islands.

Overall, the scatterplots reveal a more nuanced picture of the acceptability of *whether*-island violations than the group average. Participants did not consistently rate *whether*-island violations around the midpoint of the scale. Instead ratings were characterized by a great deal of variability at all levels. The intermediate average z-score is therefore best understood as a product of averaging over a large number of trials in which participants accepted *whether*-island violations and a smaller number of trials where participants rejected *whether*-island violations outright.

Before concluding, we would like to consider (and reject) one potential source of inconsistent ratings. Up till this point we have only considered effects and consistency on a by-subject basis, ignoring the contribution of individual items. It is logically possible that what appears to be inconsistency at the subject level was actually driven by consistency at the item-level. For example, if an individual item were unacceptable for some reason orthogonal to our *whether*-island manipulation, then participants who rated this item, but who would otherwise accept whether island violations, would erroneously appear to be inconsistent. To test whether

the inconsistency was driven by inter-item differences in acceptability, we plotted the distribution of ratings for each of the 8 items in our *whether*-island experiments in Figure 6.

Figure 6: Distributions of (z-score transformed) ratings for each whether island-violating item in Norwegian experiments 1 - 3 and the first eight whether island violating items (out of sixteen) from Sprouse et al. 2012 experiment 2.

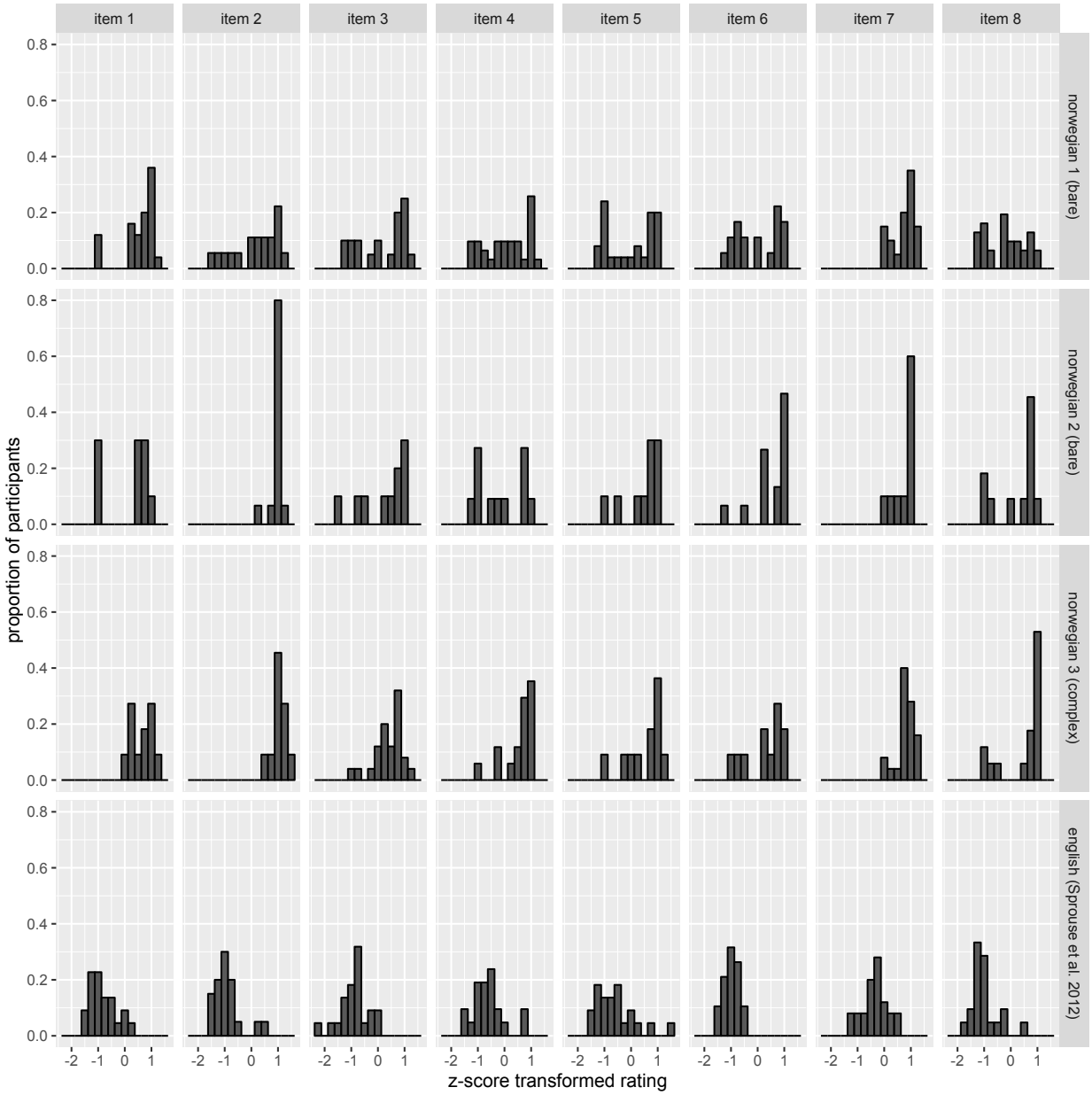


Figure 6 suggests that there was no subset of items that was uniformly responsible for the instances of relatively high ratings for *whether*-island violating sentences. Nor does it seem that inconsistency can be blamed on a group of items items that were consistently rated unacceptable. In experiment 1, all eight of the Norwegian items show distributions that suggest both acceptable and unacceptable ratings. This stands in relatively stark contrast to the English items, which show ratings that are primarily unacceptable. In experiment 1, item 7 was rated predominantly more acceptable than any of the other items. Similar rating distributions were observed for the same items in experiments 2 and 3: most items received both acceptable and unacceptable ratings. In experiments 2 and 3, most participants rated item 2 and item 7 as acceptable. We point out that the relative acceptability of items 2 and 7 alone is not sufficient to explain the number of true accepters that we saw in all three experiments. Given the construction of the lists, there were no participants whose two *whether*-island tokens were items 2 and 7. On the basis of the distributions in Figure 6, we conclude that the variability that we saw in our experiments cannot be attributed to confounds at the item level.

4. General Discussion

We investigated island effects in Norwegian using the factorial design of island effects originally explored in Sprouse (2007), Sprouse et al. (2011) and Sprouse et al. (2012) in order to better understand the range of cross-linguistic variation in island sensitivity. In particular, we were interested in determining whether we could experimentally verify claims that Mainland Scandinavian languages such as Norwegian allow filler-gap dependencies that cross embedded questions, complex NPs, and relative clauses in violation of commonly assumed universal prohibitions on such dependencies. We were also interested in pinpointing potential sources for the occasional inconsistency that has characterized judgments (particularly of complex NP and RC island phenomena) in the past literature.

We found statistically significant super-additive interaction effects for all five island types that we tested. Norwegian participants displayed adjunct and subject islands effects across all three of experiments that were comparable in size to adjunct and subject island effects in experiments in English and Italian. This result was not unexpected, as it is generally agreed that Mainland Scandinavian languages are sensitive to adjunct and subject islands. Perhaps more surprising in light of previous literature (e.g., Christensen, 1982; Allwood, 1982; Engdahl 1982, 1997), we also found clear evidence of complex NP and RC island effects in Norwegian. Complex NP and RC island effect sizes were not significantly different from the adjunct and subject island effects within our own experiments, nor significantly different from adjunct, subject, and complex NP island effects that have been tested previously in English and Italian. We return to how these effects might be reconciled with the view that Mainland Scandinavian languages are not sensitive to complex NP or RC islands after we discuss our *whether*-island findings.

Our experiments did uncover one area in which judgments in Norwegian differed from other languages that have been studied using the factorial design. We observed consistent *whether*-island effects in experiments 1 – 3, but these were roughly half the size of *whether*-island effects in English or Italian. Closer inspection of the smaller effect revealed considerable inter-individual variation in whether island sensitivity. In all three experiments, there was a substantial portion of participants (30%, 52%, 47%, respectively) that exhibited no *whether*-island effect whatsoever. In addition to these *whether*-island "accepters", there were also participants that consistently rejected *whether*-island violations in experiments 1 and 2. Thus,

rather than a consistent effect across participants, the smaller effect represented the result of averaging across groups of participants with distinct response profiles. One final – and curious – finding was that there was a non-negligible number of participants in each experiment that rated *whether*-island violations inconsistently; accepting one token that they encountered, while rejecting the other.

4.1 Meta-theoretical Implications and Open Questions

We tested a wide range of traditional islands in Norwegian and found reliable island effects in domains whose islandhood has been disputed and those whose islandhood has not. We found consistent subject and (conditional) adjunct island effects across our experiments, which indicates that traditional analyses of these two islands can be ported over to MSc languages without significant revision. Thus, it appears to us that CED-based approaches (Huang, 1982; Uriagereka, 1999; Jurka, 2010) or structure-building approaches (Uriagereka, 1999; Nunes & Uriagerka, 2000; Stepanov, 2007) to these islands are equally well supported by our results. Because our effects do not complicate – or distinguish between – the consensus views of these islands, we do not dwell on them further. We instead move on to how our results inform our understanding of *whether*-, complex NP, and RC islands in Norwegian and MSc languages more generally.

One of the goals of this paper was to winnow down the list of possible sources of the unacceptability associated with superficial island violations in MSc. Despite claims that embedded questions, complex NPs, and RCs are not syntactic islands in MSc languages, it has been consistently noted that extraction from these domains often results in unacceptability (e.g., Christensen et al. 2012, 2014; Engdahl, 1997; Maling & Zaenen, 1982). Some authors (e.g., Christensen et al. 2012, 2014) have contended that this unacceptability is not grammatical in origin, attributing it to extra-grammatical 'processing factors' such as memory load. Our results cast doubt on claims that reduce all detectable unacceptability in such constructions to simple (linearly additive) processing burdens because we found that *whether*, complex NP, and RC island effects persisted after we explicitly factored out the two most often cited processing factors (dependency length and basic structural complexity), as well as any other factors that are evenly distributed across the factorial subtraction. Our results are only compatible with either a complex processing explanation or a grammatical explanation. A number of previous studies have pointed out the challenges that face a complex processing explanation, such as the existence of cross-linguistic variation (e.g., Rizzi, 1982, Sprouse, Caponigro, Greco & Cecchetto, 2016), the existence of parasitic gaps (e.g., Engdahl, 1983, Phillips, 2006), the existence of island effects with *wh*-in-situ (e.g., Huang, 1982, Lasnik and Saito, 1992), the lack of correlation between working memory capacity and island effects (e.g., Sprouse, Wagers & Phillips 2012, Michel 2014), and the island-insensitivity of non-A' dependencies (e.g., Yoshida et al. 2015). We take the preponderance of evidence to suggest that grammatical explanations are a profitable avenue to pursue at this time, therefore we focus on this avenue in the rest of this discussion.

One of the important theoretical upshots of our studies is that different factors govern the apparent acceptability of extraction from *whether*-, complex NP, and RC islands. *Whether*-islands were the only islands to which participants in our experiments exhibited any signs of insensitivity. Our theories should reflect this fact: we must provide an explanation for *whether*-island insensitivity that separates it from all other islands on at least some (yet to be determined) dimension. One question that we cannot answer here, but which we should bear in mind when evaluating the theoretical accounts of *whether*-island insensitivity, concerns the generality of our

whether-island results: Is the degree of (variable) insensitivity that we observed specific to *whether*-islands, or should we expect the same degree of variable insensitivity to be a property of *wh*-islands on the whole? In our discussion below we consider analyses that tie insensitivity to idiosyncratic properties of Norwegian embedded polar questions headed by *om*, as well as those that extend insensitivity to all embedded questions.

Finally, our results strongly suggest that any account of *whether*-island insensitivity must countenance the fact that there is significant individual variation in absence of *whether*-island effects. We believe that a truly successful account of *whether*-island (in-)sensitivity in Norwegian should be flexible enough to tie *whether*-island sensitivity to properties of individual participants and should make explicit claims about which of its component parts (i) are subject to inter-individual grammatical variation or (ii) might be expected to be variably implemented during real-time language processing. On our view, the presence of inter-individual variation has potentially important consequences for our theories of islands, and should not be ignored. This stance has methodological implications for the growing field of experimental syntax. Experimental syntax has, to date, primarily focused on drawing inferences from differences in average acceptability calculated at the group level. Our data show that restricting attention to differences at the group level alone may cause researchers to overlook information that is theoretically relevant or to draw spurious conclusions about central tendencies in the data that do not actually exist. We would like to take this space to advocate that future work in experimental syntax provide more information about individual variation among participants in the hopes of providing a more holistic picture of the phenomena under investigation. We have offered some suggested analyses that may be useful in this regard such as plotting the distribution of DD scores, and plotting the consistency of judgments across multiple tokens of the same condition.

We now turn to more targeted discussion of how to accommodate our results within existing theories of island effects.

4.2. Whether Islands

Below we consider how our *whether*-island results could be handled within different theoretical approaches to island effects.

4.2.1. Cycle-based analyses

Cycle-based analyses of islands, which we take to encompass *Subjacency* (Chomsky, 1973, 1977), *Barriers* (Chomsky, 1986), and modern phase-based frameworks (e.g., Chomsky, 2001), hold that (some) island effects arise when long-distance A'-movement must proceed in "one fell swoop" across more than one cyclic domain. Under these analyses movement out of a finite clause must at least stopover in SpecCP (the modern-day S'). It is commonly assumed that there is only one SpecCP per finite clause, and if a finite clause's specifier is already occupied, long-distance A'-movement from that finite clause is blocked. Cycle-based analyses of *wh*-islands posit that a *wh*-operator blocks movement out of embedded questions.

One natural way to account for variation in *wh*-island sensitivity within cycle-based frameworks is to relax the assumption that there is only one specifier at the edge of a clause through which to move. Reinhart (1981) explained the apparent acceptability of *wh*-islands in Hebrew by positing that the Hebrew clause provided an extra specifier (a second COMP in Reinhart's original terminology) for successive cyclic-movement. The availability of this second COMP was presumed to vary (parametrically) across languages.

Christensen & Nyvad (2014) propose a modern variant of a multiple specifier-analysis to account for acceptable island violations in Danish (and by extension other Mainland Scandinavian languages like Norwegian). According to Christensen & Nyvad's proposal, the grammars of MSc languages allow speakers to generate multiple 'stacked' CP phrases in the left-periphery of a clause on an as-needed basis. Each of these phrases has a specifier that can serve as an intermediate landing site for successive cyclic movement. Insofar as the account can guarantee that only the top-most C in any clause is treated as the bounding node/phase head[8], the analysis makes it possible to extract from *whether*-islands and other embedded questions without violating locality.[9] Thus, the account would provide a way to explain the absence of a syntactic *whether*-island effect

There are two ways in which such a multiple-specifier analysis could accommodate inter-individual variation in *whether*-island effects. First, one might posit a grammatical difference at the population level: one group of Norwegians have grammars that allow stacked CPs and therefore permit extraction from all embedded questions, while another group does not. Although we cannot rule this analysis out completely, we consider the analysis unlikely because it does not provide a straightforward explanation for the behavioral pattern of inconsistent participants. The account predicts that individual participants should be consistent accepters or rejecters (on the assumption that participants use the same grammar across trials). Second, the multiple-specifier analysis could account for variability by supposing that all participants possess grammars that license stacked CPs, but that some participants occasionally fail to adopt a stacked CP parse for *whether*-island violating sentences. On trials in which participants did not generate the extra specifier, their parses would violate locality restrictions and an island effect would ensue. If this is the right analysis, it would seem that some of our participants adopted the correct parse reliably, while others did so probabilistically, or never at all. As before, we would still need to understand what individual-level factors dictate whether participants would successfully adopt the right parse. More importantly, we would also need to provide a rationale for why participants would fail to adopt the appropriate parse to avoid a *whether*-island violation, if their grammar makes available the multiple-specifier analysis.

4.2.2. Scope Intervention

The discussion above presupposes that our *whether*-island effects reflect a violation of some kind of cyclic bounding constraint, but it is also possible that the effects could be linked to other factors that are known to contribute to the unacceptability of extraction from embedded questions. Below we explore whether and how the effects might instead be understood as instances of *intervention* effects.

It has been reported (based on informal judgment studies) that native speakers of English often accept movement of an argument *wh*-phrase from an embedded question, but reject adjunct movement from the same domain.

(17) a. Which car did you wonder [whether to fix __]?
b. *Why did you wonder [whether to fix the car __]?
c. *How did you wonder [whether to fix the car __]?

[8] The authors are not clear on how to ensure this, though we speculate that it might be effected through a mechanism like den Dikken's *Phase Extension* (den Dikken 2007), Gallego's *Phase Sliding* (Gallego, 2010), or Bobaljik and Wurmbrand's (2005) dynamic notion of *domain*.

[9] The account was initially designed to explain the ability to extract from RCs. We return to this point later.

The same argument-adjunct asymmetry has been (informally) observed in other configurations such as Ross' (1983) Negative islands, where 'bounding' is not at issue: arguments, but not adjuncts, are easily moved across negation.

(18) a. Which car don't you think [that John fixed __]?
b. *Why don't you think [that John fixed the car __]?
c. *How didn't you wonder [whether to fix the car __]?

Many theorists treat the phenomena in (17) and (18) as (*scope*) *intervention* effects. In both cases, a scope-taking operator (*whether* in 16, *not* in 17) appears to block movement of some lower operators. Below we outline how our variable *whether*-island effects could be explained as instances of scope intervention either within a syntactic or a semantic framework.

Rizzi's (1990, 2004) Relativized Minimality (RM) represents one influential framework that explains intervention effects in syntactic terms. Roughly speaking, RM blocks a dependency between an item, A, and second item in A's c-command domain, B, if a third item, C, intervenes between A and B and C could potentially engage in a dependency with A. C is a potential dependent of A if it overlaps with B in the features that would be checked by the dependency created (see also Starke 2001). According to RM, it is impossible to successively-cyclically move a *wh*-phrase like *which tacos* across a c-commanding *whether* because both phrases are operators (they both bear the [+Op] feature). On the assumption that *om* is similarly analyzed as an operator, embedded questions headed by *om* should be islands in Norwegian, just as in English.

(19) [___ Roar wondered [whether$_{[+Op]}$ Torgeir ate which tacos$_{[+Op]}$]]

If intervention arises because *om* is an operator, one way to explain variable *whether*-island effects would be to assume that there is variation in whether *om* is analyzed as an operator ([+Op]) or a non-operator ([-Op]).

(20) [___ Roar lurer på [*om*$_{[+Op/-Op]}$ Torgeir spiste hvilke tacos$_{[+Op]}$]]

This account would explain the cross-linguistic difference in *whether*-island effects by positing that *whether* is always an operator. Some suggestive evidence that there are syntactic differences between *om* and *whether* is that *om* is not a *wh*-word (*hv*-word) in Norwegian (unlike *whether* in English). The item also functions as a preposition (20) that (unlike prepositions in English) can take a [-wh] tensed CP complement (22, as in our complex NP items). It can also function as a conditional complementizer akin to English *if* (see our conditional adjunct island items).

(21) Johnny fortalte Roar *om* Torgeir.
Johnny told Roar *about* Torgeir.

(22) Hvem rapporterte nyheten om at Anders vant medaljen?
who reported news.DEF about that Anders won medal.DEF
'Who reported the news that Anders won the medal?'

If differences in the feature composition of *om* determine the presence of *whether*-island effects, variation in island sensitivity might plausibly track whether individual participants assign *om* the [+Op] feature. Accepters would treat *om* as [-Op], whereas consistent rejecters would always assign it [+Op]. In order to explain the behavior of participants who gave inconsistent ratings, the account would have to allow individual participants to vary the feature specification of *om* across trials. One problem with such an analysis is that it seems to make the wrong predictions with respect to extraction of adjuncts from *om* clauses. If Norwegians treated *om* as a non-operator, they would be predicted to allow movement of *wh*-adjuncts from a *whether*-island as easily as *wh*-arguments. Prior literature has claimed that MSc speakers consistently judge adjunct extraction to be unacceptable (23).[10]

(23) *Hvordan lurer gjesten på [om Hanne bakte kaken t]?
How wonder guest.DEF on whether Hanne baked cake.DEF
*'How did the guest wonder if Hanne baked the cake?'

Given the purported unacceptability of (21), it would seem that *om* is always analyzed as an operator – and therefore a potential intervener – when it heads an embedded question. (We concede that this is not as strong an argument, as we did not test whether *wh*-adjunct extraction exhibits the same variation in experiments.) Finally, it should be noted that this analysis cannot generalize to explain Norwegian (purported) insensitivity to other *wh*-islands, because it is unlikely that Norwegians ever treat *wh*-phrases like *which man* as [-Op]. Again, we did not test full *wh*-islands here, so we do not know whether they show the same variation as embedded *whether*-questions.

If *om* is an intervener, how else might we explain variable *whether*-island sensitivity in terms of scope intervention in RM? RM provides one additional means of overcoming scope intervention. Rizzi (1990) – following a suggestion originally made by Cinque (1989) – proposes that the *referentiality* of a *wh*-phrase determines its ability to overcome scope intervention effects. He suggests that a *wh*-phrase that is (i) assigned an argument theta role and (ii) is *D(iscourse)-linked* bears a referential index. Following Pesetsky (1987), Rizzi treats a *wh*-phrase as *D-linked* if it was linked to a contextually salient set in the discourse representation.

Having a referential index allows a *wh*-phrase like *which man* in (24) to bind its trace across an intervener, just as the QP *every man* may bind the pronoun *him* in (25):

(24) *Which man*$_1$ did Roar wonder whether Sigrid would talk to t_1.
(25) *Every man*$_1$ wondered whether Sigrid would talk to *him*$_1$.

Rizzi argues that the possibility of binding in (25) removes the need to establish a movement chain between the *wh*-phrase and its trace. Rizzi assumes that adjuncts are not assigned referential indices, so this long-distance binding strategy is not available to them. The only way that adjunct traces can be bound is through an (antecedent-government) chain created by movement, but movement of the adjunct across *whether* is precluded by scope intervention. Thus, adjunct extraction is impossible.

If the *referential/D-linked* status of a *wh*-phrase determines whether it can overcome scope intervention effects, then an analysis of variable *whether*-island sensitivity might be based

[10] Of course, this claim merits more rigorous experimental verification so that the comparison with our results would be appropriate.

on participants' success in adopting a referential/D-linked reading of a *wh*-phrase. Since D-linking requires establishing a link between a *wh*-phrase and (set of) referent(s) in a discourse representation, the consistent accepters in our experiments would represent participants with more elaborated discourse models or participants who are more easily able to posit a relevant entity in the discourse to which to link the *wh*-phrase. Consistent rejecters would be those who have difficulty adopting a D-linked interpretation. Inconsistent raters would be participants who, for any number of reasons, failed to adopt the required reading.[11]

Recently, Rizzi (2013) has adopted a different explanation for the lack of intervention effects for complex *wh*-phrases, one he calls featural Relativized Minimality (fRM). Under the fRM approach, there is a gradient for intervention effects: the strongest intervention effects occur when there is complete identity in the features of the moved element and the intervener, weaker intervention effects occur when there is overlap but non-identity between the features. Complex *wh*-phrases have at least two features: the +Op feature and a referential feature that we can call +NP for ease of exposition. Because *whether* has +Op but not +NP, a weaker intervention effect obtains. Extending the fRM analysis to the variation that we observed in Norwegian *whether* islands would entail postulating that either that *om* sometimes loses its +Op feature as discussed above, or postulating that the bare *wh*-words in experiments 1 and 2 sometimes gain a +NP feature (for some reason).[12]

The D-linking and fRM theories do make at least one differential prediction: since D-linking is tied to the status of the *wh*-word in the discourse rather than its lexical form, it may be possible to "D-link" a bare *wh*-word using context (Pesetsky 1987). Thus the D-linking analysis might predict, in the limit, that context would 'convert' inconsistent participants to consistent accepters, whereas the fRM analysis makes no such prediction. We leave exploration of this possibility to future research.[13]

The analyses above provide a way of understanding scope intervention as a constraint on syntactic structure building. However, there are analyses that instead assume that intervention effects reflect constraints on semantic composition operations (Kiss, 1992; de Swart, 1992; Szabolcsi & Zwarts, 1993). Broadly speaking, these analyses assume that intervention effects occur when two conditions obtain: (i) an intervening scope-taker requires that a particular operation be performed in the denotation domain of the extracted *wh*-phrase and (ii) the required operation is undefined for the domain denoted by the *wh*-phrase (typically because the *wh*-phrase denotes a *partially-ordered* domain). For example, an account like Szabolcsi & Zwarts (1993) explains negative islands as follows: Negation is an operator that performs the *complementation*

[11] Miyagawa (2004), building off Beck's (1996) notion of a *Quantifier Induced Barrier*, argues that scope intervention effects emerge when a *wh*-phrase is separated from its restrictor by a scopal operator (such as *whether*). 'Referential' *wh*-phrases – which Miyagawa terms 'presuppositional' following Cresti (1995), Beck & Kim (1997), and others – avoid scope intervention because their restrictors are interpreted 'high' above the intervener. Under this implementation, variable scope intervention effects track whether participants adopt a presuppositional/non-presuppositional reading of the *wh*-phrase, because this choice determines the position of the *wh*-phrase's restrictor at LF. Under this approach, the variability that we see in Norwegian *whether* islands must correlate with whatever triggers presuppositional versus non-presuppositional readings of the *wh*-phrase.

[12] An anonymous reviewer notes that adding a +NP feature to a bare *wh*-word might be seen as reducing to a formal encoding of D-linking within the fRM framework.

[13] We would like to point out that we are aware of no published experimental evidence that context can "D-link" a bare *wh*-word: Sprouse (2007) was unable to create a D-linking effect on Superiority violations using context alone, and Villata et al. (2016) were unable to create a D-linking effect on *wh*-island violations using context alone. If this state of affairs continues, it either means that D-linking is the wrong analysis for these effects or that the contexts used in these experiments are not sufficient to induce the relevant discourse-linking.

operation. Manner adverbials such as *how* denote in partially ordered domains (e.g., free-join semilattices) that cannot be closed under complementation. Therefore, attempting to take the complement of *how*, which would be required to interpret (26), results in semantic failure.

(26) *How didn't Johnny say [that Roar fixed the car ___]?

Semantic approaches to scope intervention assume that *wh*-phrases are not subject to scope intervention if they range over an unordered domain of individuals. This is because all Boolean operations (complementation, addition, intersection, etc.) are defined over unordered sets. Thus (17a) is acceptable, because one can take the complement of the set denoted by *which car*. Importantly, as noted by Szabolcsi & Zwarts, *wh*-phrases that can range over individuals *who*, and *what*, can also range over properties. If a *wh*-phrase like *who* or *what* is interpreted as ranging over properties, then it is predicted to be sensitive to scope intervention. With this observation in hand, it is possible to provide an explanation for variable *whether*-island sensitivity among our participants: Consistent accepters always chose to interpret argument *wh*-phrases as ranging over individuals, while inconsistent accepters occasionally interpreted them as ranging over properties.[14]

4.2.3 Synthesizing the Accounts of whether-Islands and Cross-linguistic Differences

Before concluding this sub-section, we would like to make one point clear that has been implicit up till now. Following standard assumptions, we take the acceptability of *wh*-island violations in English (and similar languages) to be governed by both a bounding constraint and scope intervention. This assumption explains why English speakers often detect residual unacceptability in *wh*-island violations even if scope-intervention is ameliorated (e.g., through D-linking). If the scope intervention account of variable Norwegian *whether*-island effects is on the right track, it would appear that overcoming scope-intervention results – at least for consistent accepters – in complete acceptability. Thus, it would seem that there is no supplemental bounding constraint violation. We are inclined to interpret the fact that there were essentially no consistent rejecters in Experiment 3 as support for the hypothesis that there is relative uniformity across Norwegian speakers in this regard: whether-islands – and perhaps all *wh*-islands - are true weak islands in Norwegian. This separates Norwegian from English, but puts it on par with languages like Hungarian, about which similar claims have been made (Szabolcsi & Zwart, 1993). This therefore either suggests that Norwegians allow successive cyclic movement through the left-edge of a *whether*-island, consistent with a multiple-specifier account, or that CP is not a 'bounding node'. Where individuals differ is not in the structure that they assign to apparent *whether*-islands, but rather how easily they accommodate the relevant reading to overcome intervention effects. We leave testing this hypothesis further to future research.

4.3. Reconciling Our Findings with Prior Claims: RC Islands & CNPC

[14] We point out that this interpretation differs from the D-linking explanation outlined above in that a background context is, in principle, not necessary for adopting an individual reading of the *wh*-phrase. Szabolcsi & Zwarts (1993) do note, however, that D-linking may assist in allowing participants to generate an individual reading of an otherwise naturally ordered domain, or may speed up search through an unordered domain.

Our results appear at first blush to challenge the predominant view that complex NPs and RCs are not islands in MSc languages like Norwegian. If they were not islands, we should not have observed any super-additive effects.

Our findings complicate the picture of island sensitivity in MSc languages, but we do not wish to suggest that they invalidate previous work based on informal binary acceptability judgments. As we mentioned above, it has long been known that speakers of MSc languages reject some dependencies that span complex NPs and RCs, but not others (e.g., Taraldsen, 1979, 1982; Allwood, 1982; Engdahl, 1982, 1997; Christensen, 1982). Thus, it seems relatively uncontroversial to assert that extraction from complex NPs and RCs causes island effects. What has been a point of controversy has been whether this constraint should be stated syntactically, or whether it is better understood as non-structural (i.e. semantic, discourse-pragmatic, or even processing-based) in origin. We hope that our experiments provide a motivation and a new framework for conducting more targeted research into the factors that affect acceptability from extraction. We lay out some hypotheses that future research could explore.

One way to reconcile our results with prior claims would be to be to assume that there is a yet-unknown distributional or syntactic restriction on extraction from complex NPs and RCs in Norwegian that our items failed to satisfy. We consider this explanation somewhat unlikely given that our materials (especially our RC island items) were modeled after purportedly acceptable island violations.

A second avenue for reconciliation would be to suppose that RC island effects do not apply to all types of A'-movement uniformly. Our results show island effects for *wh*-movement in Norwegian, but we have not established that the same holds true for other long distance dependencies. We note that the majority of naturally-occurring RC-Island violations involve topicalization (Christensen, 1982; Taraldsen, 1982, see Engdahl, 1997 and Lindahl, 2014 for discussion of similar examples in Swedish). It may be the case that we would not see the same island effects in experiments that used topicalized phrases as fillers. There is some evidence in the literature that island sensitivity can vary as a function of dependency type. Sprouse et al. (2016) found that *wh*-movement out of an adjunct *if*-clause in English results in a clear super-additive island effect, but relativization out of the same structure does not; similarly, they found that *wh*-movement out of *whether* and subject islands in Italian lead to super-additive island effects, but relativization out of the same structures does not. Theories that reduce island constraints to RM effects could potentially distinguish between topicalization and *wh*-movement out of RCs. One possibility is that an RM account could posit that the relative operator acts as an intervener for *wh*-movement, but not for topicalization. If we assume that *wh*-movement and relativization are both instances of operator movement, *wh*-phrases and relative operators should both bear a generic [+Op] feature. If *wh*-movement is driven (in part) by the need to check an [Op] feature, an intervening [+Op] relative head should create a RM violation. On the other hand, the relative operator should not act as an intervener for the purposes of topicalization if topicalization is not driven by an [Op] feature. This assumption seems motivated given that topicalization is known not to exhibit some characteristics of operator movement (e.g., it is not subject to Weak Crossover, see Lasnik & Stowell, 1991; Rizzi, 1997). Another possibility is that an RM account could recognize different types of Op features (e.g., +Q, +RC, +Top, etc). One could then organize these features into classes (or hierarchies) that describe the way that they interact with each other (see Rizzi 2013 and Abels 2012 for concrete proposals along these lines). We leave exploration of these possibilities to future research.

The third option for reconciliation would be to follow authors (e.g., Erteschik-Shir, 1973; Engdahl, 1997) who tie the unacceptability associated with extractions from complex NPs and RCs to a semantic or discourse-pragmatic (SDP) – as opposed to syntactic - constraint violation. The intuition behind many SDP accounts is that participants judge extraction from complex NPs and RCs to be unacceptable when they cannot imagine or coerce a hypothetical discourse context in which the presuppositions of the island-violating structure are accommodated. The prediction of such accounts is that acceptability of extraction from complex NPs and RCs should increase if participants are given contexts that license the discourse function of the extraction and which minimize the number of accommodating assumptions that the hearer must make. Our test sentences were presented *in vacuo*, so it is possible that participants were unable to accommodate the appropriate reading of the sentence that would satisfy the relevant SDP constraint(s). In order to test whether this hypothesis is on the right track, future experiments using the factorial design should be run which pair test sentences with contexts that facilitate the appropriate reading as best they can. We note that constructing such contexts is not a trivial task because the discourse constraints that are assumed to operate on these extractions are not well understood.

5. Conclusion

We conducted an experimental survey of island phenomena in Norwegian in the hopes of better understanding whether Norwegian speakers accept violations of universal island constraints. Our studies found no evidence that naive Norwegians differed from their English (or Italian) counterparts in their sensitivity to subject, adjunct, complex NP, or RC island effects. Our complex NP and RC island results are potentially inconsistent with previous claims that Norwegian does not obey complex NP or RC islands. Interestingly, our studies uncovered one area where Norwegian judgments deviate from the cross-linguistic norm: Norwegians exhibit significant inter-individual variation in their sensitivity to *whether*-island effects. We offered some suggestions on how to investigate inter-individual variation within the context of formal acceptability judgment studies and some speculation on how such variation could inform our understanding of the grammatical basis of island effects. We hope that our work will provide new motivation and a new framework for conducting more targeted research into the factors that determine island effects in the future.

References

Abels, Klaus. 2012. The Italian Left Periphery: A View from Locality. *Linguistic Inquiry* 43: 229-254.

Adams, Marianne. 1985. Government of empty subjects in factive clausal complements. *Linguistic Inquiry* 16: 305-313.

Åfarli, Tor A., and Kristin Melum Eide. 2003. Norsk Generative Syntaks. Novus-Verlag.

Allwood, Jens. 1982. The Complex NP Constraint in Swedish. In *Readings on Unbounded Dependencies in Scandinavian Languages*, eds. Elisabet Engdahl and Eva Ejerhed, 15-32. Stockholm: Almquist & Wiksell International.

Andersson, Lars-Gunnar. 1982. What is Swedish An Exception to? Extractions and Island constraints. In *Readings on Unbounded Dependencies in Scandinavian Languages*, eds. Elisabet Engdahl and Eva Ejerhed, 33-45. Stockholm: Almquist & Wiksell International.

Beck, Sigrid, and Shin-Sook Kim. 1997. On wh- and operator scope in Korean. *Journal of East*

Asian Linguistics 6: 339-384.
Bobaljik, Jonathan D., and Susie Wurmbrand. 2005. The domain of agreement. *Natural Language and Linguistic Theory* 23: 809-865.
Boeckx, Cedric. 2008. Islands. *Language and Linguistics Compass* 2: 151-167.
Chomsky, Noam. 1964. *Current Issues in Linguistic Theory*. The Hague: Mouton.
Chomsky, Noam. 1973. Conditions on transformations. In *A Festschrift for Morris Halle*, eds. Stephen Andersen & Paul Kiparsky, 232-286. New York: Holt, Rinehart and Winston.
Chomsky, Noam. 1977. On wh-movement. In *Formal Syntax*, eds. Peter Culicover, Thomas Wasow, and Adrian Akmajian, 71-132. New York: Academic Press.
Chomsky, Noam. 1986. *Barriers*. Cambridge, MA: MIT Press.
Chomsky, Noam. 2001. Derivation by phase. In *Ken Hale: A Life in Language*, ed. Michael Kenstowicz, 1-52. Cambridge, MA: MIT Press.
Christensen, Kirsti Koch. 1982. On Multiple Filler-Gap Constructions in Norwegian. In *Readings on Unbounded Dependencies in Scandinavian Languages*, eds. Elisabet Engdahl and Eva Ejerhed, 77-98. Stockholm: Almquist & Wiksell International.
Christensen, Ken Ramshøy, and Anne Mette Nyvad 2014. On the nature of escapable relative islands. *Nordic Journal of Linguistics* 37: 29-45.
Christensen, Ken Ramshøj, Johannes Kizach, and Anne Mette Nyvad. 2012. Escape from the Island: Grammaticality and (Reduced) Acceptability of *wh*-island Violations in Danish. *Journal of Psycholinguistic Research* 42: 51-70.
Cinque, Guglielmo. 1989. On the Scope of 'Long' and 'Successive' Cylic Movement. Paper presentend at the Second Princeton Workshop on Comparative Grammar.
Cole, Peter, and Gabriella Hermon. 1994. Is there LF WH Movement? *Linguistic Inquiry* 25: 239-262.
Cresti, Diana. 1995. Extraction and Reconstruction. *Natural Language Semantics* 3: 79-122.
Deane, Paul. 1991. Limits to attention: A cognitive theory of island phenomena. *Cognitive Linguistics* 2: 1-63.
Dikken, Marcel den. 2007. Phase Extension: Contours of a theory of the role of head movement in phrasal extraction. *Theoretical Linguistics* 33: 1-41.
Drummond, Alex. 2012. *IbexFarm (Version 0.3.7) [Software]*. Available at: http://spellout.net/ibexfarm
Engdahl, Elisabet. 1982. Restrictions on Unbounded Dependencies in Swedish. In *Readings on Unbounded Dependencies in Scandinavian Languages*, eds. Elisabet Engdahl and Eva Ejerhed, 151-174. Stockholm: Almquist & Wiksell International.
Engdahl, Elisabet. 1983. Parasitic gaps. *Linguistic Inquiry* 6: 5-34.
Engdahl, Elisabet. 1997. Relative clause extractions in context. *Working Papers in Scandinavian Syntax* 60: 51–79.
Erteschik-Shir, Nomi. 1983. On the nature of island constraints. Doctoral dissertation, MIT.
Fanselow, Gisbert. 2001. Features, Θ-roles, and Free Constituent Order. *Linguistic Inquiry* 32: 405–437.
Featherston, Sam. 2005. Magnitude estimation and what it can do for your syntax: Some wh constraints in German. *Lingua* 115: 1525-1550.
Gallego, Ángel J. 2010. Phase theory and phase sliding. In *Phase Theory*, ed. Ángel J. Gallego, 51-142. Amsterdam: John Benjamins.
Goodall, Grant. 2015. The D-linking effect on extraction from islands and non-islands. *Frontiers in Psychology*, http://dx.doi.org/10.3389/fpsyg.2014.01493

Grewendorf, Günther. 1988. *Aspekte der deutschen Syntax*. Tübingen: Narr.
Haider, Hubert. 1993. *Deutsche Syntax: Generativ*. Tübingen: Narr.
Han, Chung-hye, and Jong-Bok Kim. 2004. Are there "Double Relative Clauses" in Korean? *Linguistic Inquiry* 35: 315-337.
Hofmeister, Philip & Ivan Sag. 2010. Cognitive constraints and island effects. *Language* 86: 366-415.
Hoshi, Koji. 2004. Parameterization of the external D-system in relativization. *Language, Culture, and Communication* 33: 1-50.
Huang, C.T. James. 1982. Logical relations in Chinese and the theory of grammar. Doctoral dissertation, MIT.
Ishizuka, Tomoko. 2009. CNPC violations and Possessor Raising in Japanese. Ms., UCLA.
Jurka, Johannes. 2010. The importance of being a complement: CED effects revisited. Doctoral dissertation, University of Maryland.
Kayne, Richard S. 1981. ECP Extensions. *Linguistic Inquiry* 12: 93-133.
Keller, Frank. 2000. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Doctoral dissertation, University of Edinburgh.
Kluender, Robert, and Marta Kutas. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8: 573-633.
Kush, Dave, Akira Omaki, and Norbert Hornstein. 2013. Microvariation in islands? In *Experimental Syntax and Island Effects*, eds. Jon Sprouse and Norbert Hornstein, 239-264. Cambridge: Cambridge University Press.
Lasnik, Howard, and Mamoru Saito. 1992. *Move Alpha*. Cambridge, MA: MIT Press.
Lasnik, Howard, and Timothy Stowell. 1991. Weakest crossover. *Linguistic Inquiry* 22: 687-720.
Lindahl, Filippa. 2014. Relative clauses are not always strong islands. *Working Papers in Scandinavian Syntax* 93: 1-25.
Lutz, Uli. 1996. Some notes on extraction theory. *In On Extraction and Extraposition in German*, ed. Uli Lutz and Jürgen Pafel, 1-44. Amsterdam: John Benjamins.
Maling, Joan, and Annie Zaenen. 1982. A phrase-structure account of Scandinavian extraction phenomena. In *The Nature of Syntactic Representation*, eds. Pauline Jacobson, and Geoffrey K. Pullum, 229-282. Dordrecht: Reidel.
Manzini, Rita. 1992. *Locality: A Theory and Some of its Empirical Consequences*. Cambridge, MA: MIT Press.
Michel, Daniel. 2014. Individual cognitive measures and working memory accounts of syntactic island phenomena. Doctoral dissertation. University of California, San Diego.
Miyagawa, Shigeru. 2004. The nature of weak islands. Ms., MIT.
Müller, Gereon. 1991. Beschränkungen für W-in-situ. *Groninger Arbeiten zur Germanistischen Linguistik* 34: 106-154.
Nunes, Jairo, and Juan Uriagereka. 2000. Cyclicity and extraction domains. *Syntax* 3: 20-43.
Pesetsky, David. 1987. *Wh*-in-situ: movement and unselective binding. In *The Representation of (In)definiteness*, eds. Eric Reuland, and Alice ter Meulen, 98-129. Cambridge, MA: MIT Press.
Phillips, Colin. 2006. The real-time status of island constraints. *Language* 82: 795-823.
Phillips, Colin. 2013a. On the nature of island constraints I: Language processing and reductionist accounts. In *Experimental Syntax and Island Effects*, eds. Jon Sprouse and Norbert Hornstein, 64-108. Cambridge: Cambridge University Press.
Phillips, Colin. 2013b. On the nature of island constraints II: Language learning and innateness.

In *Experimental Syntax and Island Effects*, eds. Jon Sprouse and Norbert Hornstein, 132-157. Cambridge: Cambridge University Press.
Platzack, Christer. 2000. A complement-of-N^0 account of restrictive and non-restrictive relatives. In *The Syntax of Relative Clauses*, eds. Artemis Alexiadou, Paul Law, André Meinunger, and Chris Wilder, 265–308. Amsterdam: John Benjamins.
Reinhart, Tanya. 1981. A second COMP Position. In *Theory of Markedness in Generative Grammar*, ed. Adriana Belletti, 518-557. Pisa: Scuola Normale Superiore.
Richards, Norvin. 2001. *Movement in Language*. Oxford: Oxford University Press.
Rizzi, Luigi. 1982. *Issues in Italian Syntax*. Dordrecht: Foris.
Rizzi, Luigi. 1990. *Relativized Minimality*. Cambridge, MA: MIT Press.
Rizzi, Lugi. 1997. The fine structure of the left periphery. In *Elements of Grammar*, Liliane Haegeman (ed.), 281-337. Dordrecht: Kluwer.
Rizzi, Luigi. 2004. Locality and left periphery. In *Structures and Beyond*, ed. Adriana Belletti, 223-251. Oxford: Oxford University Press.
Rizzi, Luigi. 2013. Locality. *Lingua* 130: 169-186.
Ross, John Robert. 1967. Constraints on variables in syntax. Doctoral dissertation, MIT. [Published 1986 as *Infinite syntax!* Norwood, N.J.: Ablex.]
Rouveret, Alain. 1980. Sur la notion de proposition finie: gouvernement et inversion. *Langages* 60: 61-88.
Rudin, Catherine. 1988. On multiple wh-questions and multiple wh-fronting. *Natural Language and Linguistic Theory* 6: 445-601.
Sprouse, Jon. 2007. A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge. Doctoral dissertation, University of Maryland.
Sprouse, Jon, Shin Fukuda, Hajime Ono, and Robert Kluender. 2011. Reverse island effects and the backward search for a licensor in multiple *wh*-questions. *Syntax* 14: 179-203.
Sprouse, Jon, Matt Wagers, and Colin Phillips. 2012. A test of the relation between working memory and syntactic island effects. *Language* 88: 82-124.
Sprouse, Jon, Ivano Caponigro, Ciro Greco, and Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in English and Italian. *Natural Language and Linguistic Theory* 34: 307-344.
Starke, Michal. 2001. Move dissolves into merge: a theory of locality. Doctoral dissertation, University of Geneva.
Szabolcsi, Anna, and Frans Zwarts. 1993. Weak islands and an algebraic semantics of scope taking. *Natural Language Semantics* 1: 235–284.
Taraldsen, Knut Tarald. 1982. Extraction from Relative Clauses in Norwegian. In *Readings on Unbounded Dependencies in Scandinavian Languages*, eds. Elisabet Engdahl and Eva Ejerhed, 205-221. Stockholm: Almquist & Wiksell International.
Uriagereka, Juan. 1999. Multiple spell-out. In *Working Minimalism*, ed. Samuel David Epstein, and Norbert Hornstein, 251-282. Cambridge, MA: MIT Press.
Yoshida, Masaya, Nina Kazanina, Leticia Pablos, and Patrick Sturt. 2014. On the origin of islands. *Language, Cognition and Neuroscience* 29: 761-770.
Villata, Sandra, Luigi Rizzi, and Julie Franck. 2016. Intervention effects and Relativized Minimality: New experimental evidence from graded judgments. *Lingua*
Zubizarreta, Maria Luisa. 1982. Theoretical implications of subject extraction in Portuguese. *The Linguistic Review* 2: 79-96.

Acknowledgements
This work was supported, in part, by NIH NRSA grant 5F32HD080331 to DK and NSF grants BCS-0843896 and BCS-1347115 to JS. The authors wish to thank Caroline Heycock and three anonymous reviewers for feedback that helped us improveme the paper. Versions of this work were presented at a variety of venues, including the 2015 LSA conference, NTNU, UCONN, and UMASS, where audiences provided insightful discussion. We thank Alex Drummond for creating and maintaining the IbexFarm platform. Susanna Brock, Filippa Lindahl, Ragnhild Eik, and Maria Boer Johannessen provided assistance with the materials, logistical support and helpful comments.